EFFECTIVE NEURAL ARCHITECTURES FOR CONTEXT-AWARE VENUE RECOMMENDATION

JARANA MANOTUMRUKSA

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

SCHOOL OF COMPUTING SCIENCE

COLLEGE OF SCIENCE AND ENGINEERING
UNIVERSITY OF GLASGOW

NOVEMBER 2019

Abstract

Users in Location-Based Social Networks (LBSNs), such as Yelp and Foursquare, can search for interesting venues such as restaurants and museums to visit, or share their location with their friends by making an implicit feedback (e.g. checking in at venues they have visited). The users can also leave explicit feedback on the venues they have visited by providing ratings and/or comments. Such explicit and implicit feedback by the users provide rich information about both users and venues, and thus can be leveraged to study the users' movement in urban cities, as well as enhance the quality of personalised venue recommendations. Unlike traditional recommendation systems (e.g. book and movie recommendation systems), making effective venue recommendations is more challenging because we need to take into account the users' current context (e.g. time of the day, user's current location as well as his recently visited venues).

Two common techniques that are widely used in the literature for venue recommendation systems are Matrix Factorisation (MF) and Bayesian Personalised Ranking (BPR). MF is a popular Collaborative Filtering (CF) technique that can leverage the users' explicit feedback (e.g. the numerical ratings) to predict the users' ratings on the venues and hence relevant venues can be suggested to the users based on these predicted ratings. On the other hand, BPR is a pairwise ranking-based model that can leverage implicit feedback to generate effective top-K venue recommendations. In this thesis, based upon MF and BPR models, we aim to generate effective *context-aware* venue recommendation that a user may wish to visit based on the user's historical explicit and implicit feedbacks, the user's contextual information (e.g. the user's current location and time of the day) and additional information (e.g. the geographical location of venues and users' social relationships). To achieve this goal, we need to address the following challenges: namely (**C1**) modelling the users' preferences and the characteristic of venues, (**C2**) capturing the complex structure of user-venue interactions in a Collaborative Filtering manner, (**C3**) modelling the users' short-term (*dynamic*) preferences from the sequential order of user's observed feedback as well as the contextual information associated with the successive feedback, (**C4**) generating accurate top-K venue recommendations based on the users' preferences using a pairwise ranking-based model and

(**C5**) appropriately sampling potential negative instances to train a ranking-based model.

First, to address challenge **C1**, we leverage the users' explicit feedback (e.g. their ratings and the textual content of the comments) and additional information (e.g. users' social relationships) to effectively model the users' preferences and the characteristics of venues. In particular, we propose a novel regularisation technique and a factorisation-based model that leverages the users' explicit feedback and the additional information to improve the rating prediction accuracy of the traditional MF model. Experiments conducted on a large scale rating dataset on LBSN demonstrate that the textual content of comments plays an important role in enhancing the accuracy of rating prediction.

Second, we investigate how to leverage the users' implicit feedback and additional information such as the users' social relationship and the geographical location of venues to improve the quality of top-K venue recommendations. We argue that the potential negative instances can be effectively sampled based on the social correlations between users and their friends as well as the geographical influences between the users' and venues' geographical location. In particular, to address challenges **C4** and **C5**, we propose a novel pairwise ranking-based framework for top-K venue recommendations that can incorporate multiple sources of additional information (e.g. the users' social relationship and the geographical location of venues) to effectively sample the potential negative instances. Experimental results on three large scale checkin and rating datasets from LBSNs demonstrate that the social correlations and the geographical influences play an important role to the quality of sampled negative instances and hence can improve the quality of top-K venue recommendations.

Finally, to address challenges **C2** and **C3**, we propose a framework for context-aware venue recommendations that exploits Deep Neural Network (DNN) models to effectively capture the complex structure of user-venue interactions and the users' long-term (*dynamic*) preferences from their sequential order of checkins. In particular, within the framework, we propose a novel Recurrent Neural Network (RNN) architecture that can effectively incorporate the contextual information associated with the successive implicit feedback (e.g. the time interval and the geographical distance between two successive checkins) to generate high quality context-aware venue recommendations. Experimental results on three large scale checkin and rating datasets from LBSNs demonstrate the effectiveness and robustness of our proposed framework for context-aware venue recommendations. In particular, the results demonstrate that the sequential order of users' implicit feedback can be leveraged to effectively improve the effectiveness of context-aware venue recommendation system. In addition, the time intervals and the geographical distances between two successive checkins play an important role in capturing the users' short-term preferences.

Acknowledgements

This thesis has been one of the most interesting journeys of my life. I have learned tremendously from this journey. I owe my deepest gratitude to various people for their immense support during the course of my PhD.

First and foremost, I would like to express my sincere gratitude to my supervisors, Craig Macdonald and Iadh Ounis, for their endless support and inspiration. Without their insightful advice and extensive support, this work would have not been possible.

I would also like to thank my colleagues and friends at the Terrier team and the school of computing science for collaboration and support: Anjie Fang, Graham MacDonald, Xiao Yang, Xi Wang, Ting Su, Richard McCreadie, Stuart Mackie, Romain Deveaud, Saul Vargas, Sean Moran, Jeff Dalton, Nut Limosopathan, Eugene Kharitonov, David Maxwell, Horatiu Bota, David Paule, Stewart Whiting, Fajie Yuan, Colin Wilkie, Fatma Elsafoury, and Rami S. Alkhawaldeh. I have learned a lot from them. It has been an honour and a pleasure to work with them.

I am grateful to Bowei Chen and Fabio Crestani for their thoughtful feedback and suggestions during my PhD viva, to Nick Craswell and Jaap Kamps for mentoring me in SIGIR 2017 Doctoral Consortium.

I am also thankful to my mentors at SignalAI, Dyaa Albakour and Miguel Martinez, and my mentors at Microsoft, Bhaskar Mitra, Chen Zhou and Saurabh Tiwary, for great internship experiences.

Last but not least, I am most grateful to my family, Jarean Saebang, Manee Manotumraksa, Jaranee Manotumraksa, Jarani Jaranee Manotumraksa, and my uncle, Suchart Manotumraksa, for their endless support and love throughout my life.

Table of Contents

List of Figures

List of Tables

Table 1: List of Abbreviation used in this thesis.

Abbreviation	Technical Terms
GPS	Global Positioning System
LBSN	Location-Based Social Network
CAVR	Context-Aware Venue Recommendation
CF	Collaborative Filtering technique
MF	Matrix Factorisation
BPR	Bayesian Personalised Ranking
DNN	Deep Neural Networks
MLP	Multi-Layer Perceptron
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Network-Based Social Network
GRU	Gated Recurrent Units
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
SoReg	Social Regularisation technique
CMF	Comment-based Matrix Factorisation approach
JMF	Joint Matrix Factorisation-based approach
NeuMF	Neural Matrix Factorisation
GMF	Generalised Matrix Factorisation model
GBPR	Geographical-based Bayesian Personalised Ranking
SBPR	Social-based Bayesian Personalised Ranking
SWBPR	Strong and Weak-ties Bayesian Personalised Ranking
RNN-MF	Recurrent-based Matrix Factorisation approach
BiLSTM-MF	BiLSTM-based Matrix Factorisation approach
DREAM	Dynamic REcurrent bAsket Model
TimeGRU	Time-aware Gated Recurrent Units
CGRU	Context-aware Gated Recurrent Units
LatentCross	Latent Cross-based approach
STReg	Social and Textual Regularisation technique
MFW2v	Textual Matrix Factorisation-based approach
PRFMC	Personalised Ranking Framework with Multiple sampling Criteria
DRCF	Deep Recurrent Collaborative Filtering framework
GRMF	Generalised Recurrent Matrix Factorisation model
MLRP	Multi-Layer Recurrent Perceptron model
RMF	Recurrent Matrix Factorisation
CARA	Contextual Attention Recurrent Architecture
CAG	Contextual Attention Gate
TSG	Time and Spatial Gates
CRCF	Contextual Recurrent Collaborative Filtering framework

Table 2: List of notations used in this thesis.

Notation	Meaning
R	user-venue rating matrix
C	users' checkin matrix
F	users' social link matrix
$\mathcal{U}$	set of users
$\mathcal{V}$	set of venues
$\mathcal{V}_u^+$	list of venues that user u has visited
$\mathcal{S}_u$	sequence of checkins of user u
$r_{u,i}$	rating of user u on venue i
$c_{u,i,t}$	checking of user u on venue i at time t
$s_{u,t}$	sequence of checkins of user u up to time t
lat_j, lng_i	latitude and longitude of venue i
P	the latent factors of users
p_u, ϕu_u	the latent factors of user u projected from P
Q	the latent factors of venues
q_i ϕv_i	the latent factors of venue i projected from Q
ϕt	the latent factors of time
τ	particular time step of the recurrent unit
h	hidden state
Δt	the time interval between two checkins
Δg	the geographical distance between two checkins
σ	the sigmoid activation
$tanh$	the tangent activation
$\odot$	the dot product operation
$\otimes$	the element-wise product operation
$\oplus$	the concatenation operation

Chapter 1

Introduction

1.1 Motivations

With the advance of technology such as Global Positioning System (GPS) and mobile networks in smartphones, users are increasingly interested in location-based recommendations (Narayanan and Cherukuri, 2016). An increasing number of tourists who travel to unfamiliar cities nowadays use their smartphones to find interesting places to visit or activities to do based on their current location detected by their smartphones (Yao et al., 2016). An example scenario could be a Londoner who is travelling to Glasgow and looking for a nearby local restaurant for a dinner with his family. With the emergence of Location-Based Social Networks (LBSNs) such as Yelp[1] and Foursquare[2], the users can download and install the mobile applications of various LBSNs into their smartphones. Within a given application, users are able to find various types of interesting venues to visit, share their location with their friends by making a check-in at the venue they are visiting, leave their feedback in the form of ratings as well as explicitly share their opinions by commenting/reviewing the venues they visited (Liu et al., 2013; Narayanan and Cherukuri, 2016). Large amount of feedback such as checkins, ratings and comments are generated by the users in these LBSNs. This feedback can be categorised into two types based on the users' intent when they leave the feedback, namely explicit and implicit feedback. Ratings and comments represent explicit feedback, as users explicitly specify whether they like or dislike the venues they visited. In contrast to ratings and comments, the checkin behaviour of users in LBSNs can be categorised as implicit feedback, as it is derived from the users' real-world activities (i.e. visiting venues) and does not explicitly indicate whether the users like that venues. These implicit and explicit feed-

[1] https://www.yelp.com

[2] https://foursquare.com

back provide rich information about users and venues and can be leveraged to investigate many areas including: understanding the user's movements in urban cities (Noulas et al., 2011; Ying et al., 2012), modelling the user's preferences and the characteristic of venues to enhance the quality of personalised venue recommendations (Preoţiuc-Pietro and Cohn, 2013; Liu et al., 2013), as well as provide an insight into why users visit venues, which is beneficial for businesses to find potential customers (Eravci et al., 2016).

Making effective recommendations of venues that a user may wish to visit relies on the user's feedback (e.g. his/her historical checkins) and the contextual information about the user (e.g. the user's location and time of visit). In contrast to the implicit feedback of traditional recommendation systems (e.g. item purchase feedback in Amazon[3]), making a checkin at a venue in LBSNs is a physical activity that involves various types of physical context such as the geographical influences and specific temporal characteristics. For example, users are likely to checkin at parks and museums nearby their hotels in the daytime while they might later checkin at bars in the evening. These geographical influences (e.g. visiting venues nearby their hotel) and temporal characteristics (e.g. visiting different types of venues at different times of the day) make the venue recommendation task more challenging than traditional recommendation systems. Using traditional recommendation systems that do not take various types of real-word contexts into account to generate venue recommendations may not be effective as previous works (Zhang and Chow, 2015; Zhang et al., 2015a) have shown that contexts play an important role in improving the quality of venue recommendations. As a result, Context-Aware Venue Recommendation (CAVR) has gained a lot of attention from researchers in academia and industry and has become a key functionality for users in LBSNs (Liu et al., 2013; Ying et al., 2012; Yao et al., 2016; Zhao et al., 2016; Cheng et al., 2013; Deveaud et al., 2014). Dey et al. (2001) defined context as "*any information that can be used to characterize the situation of an entity that is considered relevant to the interaction between a user and an application*". In the CAVR task, the involved entity is the user, whose context can be explicitly provided by the user (e.g. preferred time of the day) or implicitly detected by sensing devices (e.g. the user's location automatically detected from GPS in his/her smartphone).

Collaborative Filtering (CF) is a widely used technique to suggest relevant venues to users based on an assumption that similar users, who share similar preferences (i.e. visiting similar venues), are likely to visit similar venues (Koren, 2010b). Venue recommendation systems (Ma et al., 2011; Hu et al., 2014; Cheng et al., 2012) have been proposed extend Matrix Factorisation (MF), a CF-based technique that is widely used to effectively predict a user's rating on venues by exploiting explicit feedback (e.g. ratings and comments) (Koren

[3] `https://www.amazon.com`

et al., 2009), to generate venue suggestions. In particular, MF-based approaches typically aim to embed the users' preferences and the characteristics of venues within latent factors, which are combined with the dot product operator to estimate the user's preference for a given venue. Then, a ranked list of venues is generated based on the predicted user-venue rating generated by the MF model. However, explicit feedback is relatively sparse in LBSNs, which can lead to a degradation of the effectiveness of the MF-based approaches (Hu et al., 2008). In addition, in practice, users only focus on the top-K ranked list of venues (Yuan et al., 2016; Ying et al., 2016; Li et al., 2015). This implies that effective ranking-based models (e.g. learning-to-rank) that aim to generate accurate top-K venue suggestions are more useful than effective rating prediction-based models (i.e. regression models) (Rendle et al., 2009). From this point of view, MF-based approaches are not expected to perform as effectively as learning-to-rank models for the CAVR task (Shi et al., 2012). For these reasons, Bayesian Personalisation Ranking (BPR) (Rendle et al., 2009) was proposed to leverage *implicit* feedback, which is more abundant than explicit feedback and largely available in LBSNs, to generate accurate ranked lists of venues. A challenge of implicit feedback from observing checkins is that only positive feedback can be observed (i.e. we only know that users visited venues but we do not know whether they liked those venues or not), and BPR models trained on only positive feedback are likely to be biased to positive instances (Rendle et al., 2009; Zhao et al., 2014; Yuan et al., 2016). To address this challenge, various negative sampling approaches have been proposed (Rendle et al., 2009; Yuan et al., 2016; Zhao et al., 2014; Wang et al., 2016). For example, the negative sampling approach proposed in BPR by Rendle et al. (2009) uniformly and randomly selects venues that the users have not visited as negative instances. Moreover, users' preferences extracted from *implicit* feedback are not *static* and change dynamically over time (e.g. users may prefer to visit shopping malls during the daytime but prefer to visit bars in the evenings) (Koren, 2010a). Unfortunately, traditional BPR models (e.g. (Rendle et al., 2009; Wang et al., 2016; Zhao et al., 2014; Yuan et al., 2016)) can only capture the users' long-term (*static*) preferences and not their short-term (*dynamic*) preferences. Since the traditional BPR models assume that the users' preferences are static, users who have visited similar sets of venues in different orders would get similar venue suggestions (Yu et al., 2016; Zhang et al., 2014b). However, the previous literature have shown that recent observed feedback can have more influence on users' behaviour than historical feedback (Zhu et al., 2017; Liu et al., 2016c; Cheng et al., 2013; Rendle et al., 2010; Yu et al., 2016). For instance, consider a user who has recently visited several art museums and a restaurant, sequentially. Models that only capture the user's long-term preferences will recommend other museums to visit, whereas a model that can capture the user's short-term preferences might recommend a bar to visit instead. In this thesis, we argue that the sequential properties of observed checkin feedback can be leveraged to effectively capture the users' *static* and *dynamic* preferences.

Another challenge of CAVR is the problem of cold-start users (i.e. users who have typically only visited and checked in a very small number of all venues in the LBSNs). The existing literature have shown that traditional BPR models typically suffer from the cold-start user problem, which hinders the quality of the personalised venue suggestions. To mitigate the cold-start user problem, various approaches have been previously proposed to leverage additional information such as social information (Zhao et al., 2014; Wang et al., 2016; Ma et al., 2011), temporal influence (Gao et al., 2013), textual content of comments (Zhang et al., 2015a) as well as geographical information (Cheng et al., 2012; Yuan et al., 2016; Zhang and Chow, 2015; Lian et al., 2014; Ying et al., 2016; Li et al., 2015). In particular, a common approach that enhances the performance of BPR models under cold-start conditions is to extend the sampling criterion and pairwise ranking function of BPR to incorporate *additional* sources of information (e.g. social links (Wang et al., 2016; Zhao et al., 2014) and the geographical information of venues (Yuan et al., 2016)). However, the sampling criterion and pairwise ranking function of BPR do not take the sequential properties of observed feedback into account and still assume that the users' preferences are static.

To leverage the sequential properties of implicit feedback in order to capture the users' *dynamic* preferences, existing approaches in the literature (e.g. (Cheng et al., 2013; Rendle et al., 2010)) have been proposed based on Markov Chains. However, such Markov Chains-based approaches have a well-known limitation in that they can only model local sequential behaviour between each pair of adjacent observed feedback (Yu et al., 2016). To effectively capture the users' *dynamic* preferences from their observed implicit feedback, various approaches (Tang et al., 2017; Yu et al., 2016; Zhang et al., 2014b) have been proposed to exploit Recurrent Neural Network models (RNN) for recommendation systems. There are several limitations of these existing RNN-based approaches. First, they still rely on the dot product of latent factors of users and venues to estimate the user's preference for a given venue. However, He et al. (2017) argued that the dot product of latent factors may not be sufficient to capture the complex structure of user-venue interactions. To address this challenge, they proposed a Neural Matrix Factorisation (NeuMF) framework that replaces the dot product operation with a neural architecture that can learn an arbitrary function from user-venue interactions. However, the NeuMF framework is not sufficiently flexible to incorporate the sequential properties of observed implicit feedback. Second, these RNN-based approaches are not suitable for CAVR because they can only take the sequential order of checkins into account and ignore the contextual information associated with the sequences of checkins (e.g. time interval and geographical distance between two successive checkins). Indeed, such contextual information have been shown to play an important role in generating effective CAVR systems (Beutel et al., 2018; Liu et al., 2016b,c; Zhu et al., 2017; Smirnova and Vasile, 2017).

1.2 Thesis Statement

The statement of this thesis is that the quality of the personalised ranked list of venues based on the user's preferred context, Context-Aware Venue Recommendation, can be effectively enhanced by leveraging the additional information such as the users' social relationships, the textual content of comments and the geographical information of venues, the sequential properties of the users' implicit feedback and the contextual information associated with the sequences of users' implicit feedback, which can be achieved by a framework that consists of the following four functionalities/components, namely (1) capturing the complex structure of user-venue interactions in a collaborative filtering manner using an effective neural architecture to learn an arbitrary function from the user's implicit feedback, (2) modelling the users' long- (*static*) and short-term (*dynamic*) preferences from the sequential order of user's checkins and the contextual information associated with the successive checkins, (3) generating accurate top-K venue suggestions based on the user's *static* and *dynamic* preferences using a pairwise ranking function and (4) sampling potential negative instances that take into account the additional information such as the geographical information of venues, the users' social relationships and the sequential order of users' checkins.

1.3 Contributions

The main contributions of this thesis are the following. Our first contribution is to enhance the effectiveness of Collaborative Filtering-based (CF) approaches such as the traditional Matrix Factorisation (MF) and Bayesian Personalised Ranking models. In particular, we propose a novel regularisation technique and a factorisation-based model that leverage the additional information (e.g. the users' social relationships and the textual content of comments associated with users' ratings) to improve the rating prediction accuracy of the traditional MF model. Then, we propose a Personalised Ranking Framework with Multiple sampling Criteria (PRFMC), an extension of the traditional BPR model, that can incorporate multiple sources of additional information during the negative sampling and ranking process. The summary of our first contribution, which consists of two sub-contributions, is described below:

- **Enhanced MF-based Approaches**: We propose a Social-Textual Regularisation (STReg) technique that leverages the users' social links and the textual content of comments associated with the users' ratings to enhance the effectiveness of the traditional MF model. In particular, the STReg technique exploits word embeddings to estimate a semantic similarity of friends based on their textual comments to regularise the complexity of the traditional MF model. Moreover, we propose a novel textual factorisation-based model (MFw2v), an extension of the traditional MF model, which exploits word embeddings to effectively model users' preferences and the characteristics of venues from the textual content of comments left by the users. Experiments conducted on a large user-venue rating dataset from a commercial LBSN demonstrate that the textual content of comments play an important role in enhancing the effectiveness of traditional MF model. In addition, our experimental results demonstrate that our proposed STReg technique and MFw2v model can outperform various state-of-the-art rating prediction approaches.

- **Enhanced BPR-based framework**: We propose a novel Personalised pairwise Ranking Framework with Multiple sampling Criteria (PRFMC) that can leverage multiple sources of additional information to enhance the quality of venue recommendation of the traditional BPR model. In particular, the PRFMC framework exploits state-of-the-art geographical and social probabilistic models that can effectively capture the users' geographical movements and social influences to effectively sample negative instances during the training process. We empirically evaluate the effectiveness of the PRFMC framework in comparison with various existing BPR-based models on three public large-scale datasets from commercial LBSNs. Our comprehensive experiments demonstrate the effectiveness of the PRFMC framework, which are superior to the current state-of-the-art BPR models.

Our second contribution is to propose a CF-based framework for sequential-based top-K venue recommendations. Firstly, we propose a Deep Recurrent Collaborative Filtering framework (DRCF) that leverages the sequential order of users' checkins to capture the users' *static* and *dynamic* preferences to effectively generate the ranked list of personalised venues. The DRCF framework consists of two components, namely (1) the neural architecture that models the complex structure of user-venue interactions in a Collaborative Filtering manner and (2) the pairwise ranking function and dynamic geo-based negative sampling approach that aim to enhance the quality of sequential-based venue recommendations as well as alleviate the problem of cold-start users. A summary of our second contribution, which consists of two sub-contributions, is described below:

- **Neural Network Architecture**: Within the DRCF framework, we propose the Generalised Recurrent Matrix Factorisation (GRMF), Multi-Level Recurrent Perceptron (MLRP) and Recurrent Matrix Factorisation (RMF) models that exploit Deep Neural Network models to capture the complex structure of user-venue interactions from observed checkins in a Collaborative Filtering manner. In particular, the GRMF, MLRP and RMF models use the element-wise, concatenation and dot product operations, respectively, to combine the latent factors of users and venues. Then, these combinations of latent factors are weighted by using a Deep Neural Network architecture. We evaluate the effectiveness of our proposed models in capturing the complex structure of user-venue interactions in comparison with various MF-based Deep Neural Network approaches previously proposed in the literature (He et al., 2017; Yu et al., 2016). In particular, our comprehensive experiments on three public large-scale datasets from commercial LBSNs demonstrate the effectiveness of the DRCF framework, which are superior to the current state-of-the-art Deep Neural Network approaches.

- **Pairwise Ranking Function & Negative Sampling Approach**: Within the DRCF framework, we propose a novel *dynamic* geo-based negative sampling approach that takes the sequential order of users' checkins as well as the geographical information of venues into account to effectively sample negative venues, which are used in the pairwise loss function. In particular, the pairwise ranking function aims to rank venues that the users have previously visited higher than the venues they have not visited before, while our proposed *dynamic* geo-based negative sampling approach aims to sample venues that users have never visited before but are located nearby the venues they previously visited in their sequence of checkins. We conduct experiments to investigate the effectiveness of our proposed ranking function and our negative sampling approach in enhancing the quality of personalised ranked list of venues and alleviating the problem of cold-start users. The experimental results show that our proposed *dynamic* geo-based negative sampling approach can significantly improve the effectiveness of the DRCF framework as well as alleviate the cold-start problem.

Next, our last contribution is to propose a CF-based framework for context-aware sequential-based top-K venue recommendations. We propose a Contextual Attention Recurrent Architecture (CARA) that effectively captures the users' *dynamic* preferences from the sequential properties of users' checkins and the contextual information associated with the successive checkins. Finally, we propose a Contextual Recurrent Collaborative Filtering framework (CRCF) that aims to generate effective ranked lists of venues to the users based on their preferences as well as their preferred context. In particular, the CRCF framework is built upon both the DRCF framework and the CARA architecture for CAVR. A summary of our third contribution is described below:

- **Contextual Attention Recurrent Architecture**: We propose a novel Contextual Attention Recurrent Architecture (CARA) for CAVR that leverages both sequences of user's observed feedback and the contextual information associated with these sequences to capture the users' *static* and *dynamic* preferences. Within the CARA architecture, inspired by the gating mechanism of Gated Recurrent Units (GRU), we propose two refined gating mechanisms: a Contextual Attention Gate (CAG) and a Time- and Spatial-based Gate (TSG). The CAG controls the influence of the user's preferred context and previous visited venues, while TSG controls the influence of the hidden state of the previous RNN unit based on the time interval and geographical distances between two successive checkins. In particular, TSG assumes that the shorter the time interval and geographical distance between two successive checkins, the more likely the previous checkin influences the users' *dynamic* preferences on the next visit venues. To the best of our knowledge, CARA is the first RNN architecture that can incorporate multiple types of contextual information associated with the successive checkins. Our thorough experiments on three large checkin and rating datasets from commercial LBSNs demonstrate the effectiveness of our proposed CARA architecture by significantly outperforming various state-of-the-art matrix factorisation approaches and existing RNN architectures, respectively.

- **Contextual Recurrent Collaborative Filtering Framework**: We propose a Contextual Recurrent Collaborative Filtering framework (CRCF) that combines the DRCF framework and the CARA architecture for CAVR. In particular, we first modify the GRMF, MLRP and RMF models of the DRCF framework to effectively capture the users' *dynamic* and contextual preferences from their sequence of checkins by exploiting the CARA architecture. Then, we modify the pairwise ranking function of the DRCF framework to take the users' preferred context into account. We evaluate the effectiveness of our proposed CRCF framework in comparison with the state-of-the-art CAVR systems on three large-scale datasets from commercial LBSNs. Experimental results demonstrate that CRCF can significantly and consistently outperform both the DRCF framework and the CARA architecture as well as the state-of-the-art CAVR systems across the three used datasets.

1.4 Origins of Material

Most of the material presented in this thesis has been published in various international conferences during the course of the PhD programme. The following lists various publications that form the basis of research detailed in the following chapters:

- Chapter 4: We propose the Social and Textual Regularisation (STReg) technique and the textual MF-based model (MFw2v) that both exploit word embeddings to effectively model users' preferences and the characteristics of venues from the textual content of comments on LBSN. These works were published at CIKM 2016 (Manotumruksa et al., 2016) and ECIR 2017 (**?**). In addition, we propose Personalised pairwise Ranking Framework with Multiple sampling Criteria (PRFMC) for venue recommendation that exploits probabilistic models to effectively sample negative examples and generate personalised venues to users. This work was published in CIKM 2017 (Manotumruksa et al., 2017b)

- Chapter 5: We propose the Deep Recurrent Collaborative Filtering framework (DRCF) for venue recommendation that exploits a Deep Neural Architecture to effectively capture the users' *static* and *dynamic* preferences from their sequence of checkins. We also propose the novel *dynamic* geo-based negative sampling and the pairwise ranking function that take the sequential order of users' checkins as well as the geographical information of venues into account. These approaches were first published at CIKM 2017 (Manotumruksa et al., 2017a). These works were also presented at the Doctoral Consortium of SIGIR 2017 (Manotumruksa, 2017).

- Chapter 6: We propose the Contextual Attention Recurrent Architecture (CARA), which leverages the sequences of users' checkins and contextual information associated with the successive checkins to capture the users' *dynamic* preferences. This architecture was first published in SIGIR 2018 (Manotumruksa et al., 2018).

- Chapter 7: We propose the Contextual Recurrent Collaborative Filtering framework (CRCF) for context-aware venue recommendation, which integrates both the CARA architecture into the CRCF framework. This framework is in press for a special issue on Deep Learning for Information Retrieval in the Information Processing and Management (IPM) journal.

1.5 Thesis Outline

The remainder of this thesis is organised as follows:

- Chapter 2 introduces important collaborative filtering concepts for Recommendation Systems (RS) that are used all throughout the thesis. In particular, we begin by describing two different types of methods that are commonly used in Collaborative Filtering, which are referred to as memory-based methods and model-based methods. Moreover, we describe types of users' feedback in LBSNs and the evaluation of recommendation systems. Furthermore, we discuss more advanced research areas in Deep Neural Network (DNN) that are used or extended in the later chapters of this thesis to tackle specific Context-Aware Venue Recommendation (CAVR) tasks.

- Chapter 3 discusses related work in venue recommendation systems previously proposed in the literature. In particular, we review various existing MF-based approaches that leverage additional information (e.g. the users' social links and textual content of comments associated with the users' ratings) to enhance the prediction accuracy of the traditional MF models. Then, we describe the state-of-the-art Matrix Factorisation framework that exploits Deep Neural Network (DNN) to model the user-venue interactions in a Collaborative Filtering manner. Moreover, we review the existing negative sampling approaches and ranking functions that aim to enhance the effectiveness of Bayesian Personalised Ranking (BPR) for top-K venue recommendations. Furthermore, we review existing RNN-based factorisation approaches and extensions of RNN architectures that capture the users' short-term (*dynamic*) preferences from the sequential order of users' checkins. Finally, we identify the knowledge gap in these existing approaches that this thesis aims to address.

- Chapter 4 describes our proposed Social and Textual Regularisation (STReg) technique and our textual MF-based model (MFw2v) that exploit word embeddings to model the users' preferences and the characteristic of venues from the users' textual comments. In addition, we describe our proposed Personalised Ranking Framework with Multiple sampling Criteria (PRFMC), which can leverage multiple sources of additional information to generate top-K venue recommendations. Then, we investigate the effectiveness of our proposed STReg technique and MFw2v model in enhancing the prediction accuracy of the traditional MF model in comparison with existing MF-based approaches previously proposed in the literature. In addition, we evaluate the effectiveness of our proposed PRFMC framework for top-K venue recommendations in comparison with the existing state-of-the-art venue recommendation systems (e.g.

He et al. (2017); Yuan et al. (2016); Yu et al. (2016)) on three large-scale checkin and rating datasets from commercial LBSNs.

- Chapter 5 describes our proposed Deep Recurrent Collaborative Filtering (DRCF) framework that exploits the Deep Neural Network (DNN) architectures to capture the complex structure of user-venue interactions from their checkin feedback. Our proposed framework consists of three components, namely Generalised Recurrent Matrix Factorisation (GRMF), Multi-Level Recurrent Perceptron (MLRP) and Recurrent Matrix Factorisation (RMF) models. In particular, GRMF, MLRP and RMF use the element-wise, concatenation and dot product operations, respectively, to combine the latent factors of users and venues. We postulate that using these different operations is more effective in capturing the user-venue interactions than the dot product operation, which is widely used in MF-based approaches. Moreover, we describe our proposed pairwise ranking function and *dynamic* geo-based negative sampling approach that takes the sequential order of users' checkins as well as the geographical information of venues into account to effectively sample negative instances. We hypothesise that venues that have not been visited by users but are nearby to venues that the users previously visited are more likely to attract the user's preferences. Then, we investigate the effectiveness of the DRCF framework and its components in comparison with various matrix factorisation approaches on three large-scale checkin and rating datasets from commercial LBSNs. Furthermore, we empirically evaluate whether our proposed ranking function and our negative sampling approach can enhance the effectiveness of the DRCF framework in producing a higher quality personalised ranked list of venues as well as alleviating the problem of cold-start users.

- Chapter 6 describes our proposed Contextual Attention Recurrent Architecture (CARA), an extension of the Recurrent Neural Network models, that effectively captures the users' *dynamic* preferences by taking the contextual information associated with the sequences of users' checkins into account. The proposed architecture consists of two gating mechanisms that aim to control the impact of context associated with the successive checkins that influences the users' *dynamic* preferences. Then, we evaluate the effectiveness of our proposed architecture in capturing the users' *dynamic* preferences and alleviating the cold-start user problem by comparing with various state-of-the-art RNN architectures (e.g. (Zhu et al., 2017; Smirnova and Vasile, 2017; Beutel et al., 2018)).

- Chapter 7 describes our proposed Contextual Recurrent Collaborative Filtering (CRCF) framework, which is built upon both our proposed Deep Recurrent Collaborative Filtering framework (DRCF) and Contextual Attention Recurrent Architecture (CARA) for CAVR. In particular, we propose to integrate the CARA architecture into the DRCF framework to effectively capture the users' *dynamic* preferences from their sequences of checkins. Finally, we investigate the effectiveness and robustness of the CRCF framework in comparison with various matrix factorisation approaches on three large-scale checkin and rating datasets from commercial LBSNs.

- Chapter 8 closes this thesis by highlighting the contributions and the conclusions of each chapter. We also discuss some possible future directions for our research.

Chapter 2

Background

In this chapter, we provide an overview of the recommendation systems in general and basic architectures of deep neural networks that this thesis builds on. Indeed, we describe the existing work that our proposed framework relies on as a basis for modelling the complex structure of user-venue interactions in a collaborative filtering manner, generating accurate top-K venue suggestions using a pairwise ranking function, sampling potential negative instances and modelling sequential properties from the users' observed sequence of checkins using deep neural networks. In particular, we first describe the general background of recommendation systems including the goals and formulations of recommendation systems in Section 2.1. Then, we provide basic models of collaborative filtering algorithms such as Matrix Factorisation (MF) and Bayesian Personalised Ranking (BPR) widely used in the literature (Rendle et al., 2009; Koren et al., 2009) in Section 2.1.1. Although these CF-based models (i.e. MF and BPR) were not originally proposed for venue recommendations, they are sufficiently flexible to be applied to the venue recommendation task. From now on, we explain these models in the context of venue recommendation for reasons of uniformity. The evaluation methodology and metrics used in recommendation systems are provided in Section 2.1.2. Finally, in Section 2.2, we describe basic architectures of deep neural networks such as Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), which are used by advanced recommendation approaches, including those proposed in this thesis.

2.1 Overview of Recommendation Systems

Recommendation Systems are software tools and techniques that provide personalised suggestions for users about a catalogue of items such as products, video, songs, venues, or other resources (Resnick and Varian, 1997; Ricci et al., 2015b; Aggarwal et al., 2016). In scenarios where there is an overwhelming amount of information, the recommendation systems aim to help the users to explore and select information based on their preferences. Personalised suggestions generated by the recommendation systems have been shown to be effective in enhancing users' satisfaction and improving the revenues of many e-commerce and media streaming platforms such as Amazon, Netflix, Youtube or Spotify as well as commercial LBSNs such as Foursquare and Yelp (Aggarwal et al., 2016; Yang et al., 2015; Narayanan and Cherukuri, 2016). In general, recommendation systems are built based on a set of multidisciplinary theories including technologies and algorithms from various fields such as Information Retrieval, Machine Learning, Human Computer Interaction, Marketing, Economics and many others (Vargas, 2015). Recommendation Systems are an interesting research topic that has attached tremendous amount of attention from both industries and academics in the last decade.

With the emergence of e-commerce and Location-Based Social Networks (LBSN) platforms, users can explicitly or implicitly provide feedback whether they like or dislike services/venues provided by the platforms. User feedback can be categorised into two categories: namely *explicit* and *implicit* feedback. With *explicit* feedback, the users consciously and explicitly express their preferences for the service/venues in the form of preference assessments such as a simple binary feedback (i.e. like or dislike), numerical ratings (typically 1-5 starts) as well as textual content of comments. In contrast, with *implicit* feedback (e.g. click, view and checkin), the users' preferences in the services/venues are indirectly observed and estimated by other variables that do not require the active involvement of the users. Indeed, the type of users' feedback largely determines the choice of algorithms and evaluation methodologies for the development of recommendation systems. Next, we describe several ways in which the recommendation systems can be formulated. The two primary approaches are described as follows:

- Rating prediction-based approaches: These approaches aim to predict the numerical rating score (e.g. 1-5 scale rating stars) of a user for a given venue. They require training data that indicates user-venue ratings to model the users' preferences for venues. Traditionally, the user-venue ratings are represented as a matrix $R^{m \times n}$ where m and n are the number of users and venues, respectively. $r_{i,j}$ indicates the observed 1-5 scale rating feedback by user i on venue j. These ratings are used for training, while the missing or unobserved values in the user-venue rating matrix R can be predicted using the trained model. This problem can be referred to as the matrix completion problem because we have an incompletely specified matrix of values, and the remaining values are predicted by the trained model.

- Ranking-based approaches: In practice, users only focus on the top-K recommendations generated by the e-commerce and social networks platforms. Unlike the rating prediction-based approaches, the ranking-based approaches are designed to directly optimise for ranking venues (i.e. focusing on getting the top-ranked venue suggestions that are relevant to users). In general, the ranked list of venue suggestions can be obtained based on the predicted user-venue rating scores generated by the rating prediction-based approaches. However, previous works (Rendle et al., 2009; Shi et al., 2012) have shown that effective ranking-based approaches that aim to generate accurate top-K venue suggestions are more useful than effective rating prediction-based approaches. We will further discuss this aspect through several empirical studies conducted in Chapter 4

2.1.1 Basic Models of Recommendation Systems

The basic models for venue recommendation systems generally leverage two types of data to generate effective venue suggestions to the users, which are (1) the user-venue interactions (i.e. user-venue ratings or checkins) and (2) the information about the users and venues such as the textual content of comments of venues left by the users. Approaches that use the former are referred to as Collaborative Filtering-based recommendation systems whereas approaches that use the latter are referred to as content-based recommendation systems (Aggarwal et al., 2016).

2.1.1.1 Collaborative Filtering-Based Recommendation Systems

Collaborative Filtering-based (CF) recommendation systems use the collaborative power to make recommendations, by leveraging the observed user-venue interactions, such as the ratings and checkins of multiple users. The main challenge in designing CF-based approaches is that the user-venue rating and checkin matrix is generally sparse, because the users only explicitly rate or checkin a small fraction of the large universe of available venues on LBSNs (i.e. most of the rating and checkin values in the matrix are unobserved). The basic idea of the CF-based approaches is that these unobserved ratings/checkins can be estimated based on the observed ratings/checkins, which are likely to be correlated across various users and venues (Aggarwal et al., 2016). In particular, CF-based approaches assume that users who share similar preferences (e.g. rating positively or negatively the same venues or checking in at similar venues) are likely to prefer similar venues (Koren et al., 2009). From the technical perspective, CF-based approaches can be categorised into *Memory-based* (Linden et al., 2003) and *Model-based* (Koren et al., 2009) approaches.

2.1.1.1.1 Memory-based Collaborative Filtering Approaches *Memory-based* CF approaches are the earliest collaborative filtering algorithms (Sarwar et al., 2001; Linden et al., 2003), which are widely used to predict the user-venue ratings. The memory-based approaches can be categorised into user- and item-based collaborative filtering approaches. *User-based* CF approaches leverage the ratings of like-minded users of a target user to predict the rating of the target venue for the target user. The basic idea is to determine the like-minded users who have similar preferences to the target user and predict the rating of the target venue by computing weighted averages of the ratings of the target venue observed from those like-minded users. For instance, users A and B have similarly rated venues in the past, then one can use A's observed rating on a target venue to predict B's unobserved rating on the target venue. In contrast, *item-based* CF approaches first determine a set of venues that are most similar to a target venue. Then, the ratings of the set of venues, which are provided by the target user, are used to predict whether the target user prefers the target venue or not. Intuitively, the user's rating on a target restaurant can be estimated based on the ratings of other restaurants specified by the user. Both *User-* and *Item-based* CF approaches (e.g. Linden et al. (2003); Sheugh and Alizadeh (2015)) rely on similarity metrics such as cosine similarity and Pearson correlation, which have been widely used to determine similar users and venues, respectively. For example, in the User-based CF approaches, we can estimate the cosine similarity between two users, A and B, as follows:

$$similarity(a,b) = cos(a,b) = \frac{ab}{\parallel a \parallel \parallel b \parallel} \tag{2.1}$$

where a and b correspond to rows the from user-venue rating matrix of user A and B, respectively. Then the predicted rating score of target venue i for target user u, $\hat{r}_{u,i}$, can be estimated as follows:

$$\hat{r}_{u,i} = \frac{\sum_{j \in P_u(i)} similarity(v_u, v_j) \cdot r_{j,i}}{\sum_{j \in P_u(i)} \mid similarity(v_u, v_j) \mid} \tag{2.2}$$

where $P_u(i)$ denotes the set of the top-k similar users of target user u, for which ratings of venue i have been observed and v_u and v_j are the corresponding rows from the user-venue rating matrix of user u and j, respectively. The advantages of *Memory-based* CF approaches are that they are simple to implement and the resulting recommendations are often easy to explain. However, the effectiveness of such approaches can be markedly degraded when the user-venue rating matrix is sparse. An example scenario can be if there are no users whose preferences are similar to the target user and have rated a target venue, then such an approach can fail to predict the rating of the target venue for the target user.

2.1.1.1.2 Model-based Collaborative Filtering Approaches Unlike the *Memory-based* CF approaches, the *Model-based* CF approaches exploit machine learning techniques to automatically learn parameters in order to capture the users' preferences from the observed user-venue interactions. Well known and effective example of the *Model-based* CF approaches is Matrix Factorisation (MF), proposed by Koren et al. (2009). Traditional MF-based approaches assume that users can influence each other if they share similar preferences, i.e. who rate venues similarly. In particular, the MF-based approaches aim to approximate the matrix R by finding a decomposition of R into two lower dimensional matrices, namely the latent factors of users $P \in \mathbb{R}^{m \times d}$ and venues $Q \in \mathbb{R}^{n \times d}$ where d is the number of latent dimensions, such that the predicted rating of user u on venue i can be computed as follows:

$$\hat{r}_{u,i} = p_u^T q_i = p_u \circledcirc q_i = \sum_{k=1}^{d} p_{u,k} \dot{q}_{i,k} \tag{2.3}$$

where $\circledcirc$ denotes the dot product and p_u and q_i are the latent factors of user u and venue i, respectively. Indeed, MF behaves as a linear model of latent factors by assuming that each dimension of the latent factor is independent and linearly combining those dimensions with the same weight (He et al., 2017). The objective of MF is to minimise the pointwise loss between the predicted rating $\hat{r}_{u,i}$ and the observed rating $r_{u,i}$ and hence the loss function of MF is defined as follows:

$$\mathcal{L}(\Theta) = \underset{\Theta}{min} \frac{1}{2} \sum_{u=1}^{m} \sum_{i=1}^{n} I_{u,i} \cdot (r_{u,i} - \hat{r}_{u,i})^2 + \frac{\lambda}{2} \|\Theta\|_F^2 \tag{2.4}$$

where $I_{u,i}$ is an indicator variable that is 1 if user u leaves a rating at venue i, otherwise 0. To avoid overfitting, i.e. $P^T Q = R$, a traditional regularisation technique is added into Equation (2.4), where $\lambda \geq 0$ is a regularisation parameter, $\Theta = \{P, Q\}$ denotes all the parameters to be learnt and $\|A\|_F^2 = \sqrt{\sum_{i=1}^{m} \sum_{j=i}^{n} |a_{i,j}|^2}$ denotes the Frobenius norm of matrix $A \in \mathbb{R}^{m \times n}$. Next, Stochastic Gradient Descent (SGD) is applied to find a local minimum of the loss function, by optimising each of the latent factors, p_u and q_i, while fixing the other, until convergence.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_u} &= \sum_{j=1}^{n} I_{i,j}(r_{i,j} - p_u^T q_i) q_i + \lambda p_u \\ \frac{\partial \mathcal{L}}{\partial q_i} &= \sum_{i=1}^{m} I_{i,j}(r_{i,j} - p_u^T q_i) p_u + \lambda q_i \end{aligned} \tag{2.5}$$

Previous works by Cheng et al. (2012); Lee and Seung (1999); Griesner et al. (2015); Lian et al. (2014) have shown that the traditional MF model can accurately predict the users' ratings on LBSNs. To generate a ranked list of venues to the users, a common approach is to rank the venues based on the predicted ratings generated by the MF model. In the next section, we discuss why the traditional MF model is not effective for venue recommendations and then describe an extension of the MF model that focuses on generating effective ranked list of venues.

2.1.1.1.3 Bayesian Personalised Ranking (BPR) In practice, as mentioned in Section 1.1, users only focus on the top-K ranked list of venues, hence effective ranking-based models that aim to generate accurate top-K venue suggestions are more useful than effective rating prediction-based models (i.e. MF models) (Rendle et al., 2009). From this point of view, MF-based approaches are not expected to perform as effectively as learning-to-rank models for the venue recommendation task (Shi et al., 2012). In addition, explicit feedback is relatively sparse in LBSNs, which can degrade the effectiveness of the MF-based approaches (Hu et al., 2008). To address these challenges, various ranking-based approaches (e.g. (Rendle et al., 2009)) have been proposed to leverage implicit feedback (e.g. checkins), which is more abundant than explicit feedback (Eravci et al., 2016; Preoţiuc-Pietro and Cohn, 2013; Yu et al., 2014; Li et al., 2016), to generate accurate venue suggestions. In particular, Rendle et al. (2009) proposed Bayesian Personalised Ranking (BPR), a popular pairwise ranking-based approach that is widely implemented and extended to leverage implicit feedback to generate the top-K venue recommendations (e.g. (Yuan et al., 2016; Wang et al., 2016; Zhao et al., 2014; Loni et al., 2016)). Using the venue recommendation terminology, the pairwise ranking criterion of BPR assumes that a user prefers visited venues observed from their historical checkins over the non-visited ones. In particular, for each user $u \in \mathcal{U}$, the likelihood function of BPR, which aims to maximise the probability that user u will checkin at venue i ($\hat{c}_{u,i}$) is higher than the probability that user u will checkin at venue j ($\hat{c}_{u,j}$), can be expressed as follows:

$$\mathcal{L}(\Theta) = \prod_{u \in \mathcal{U}} \prod_{i \in \mathcal{V}_u^+} \prod_{j \in \mathcal{V}_u^-} P(\hat{c}_{u,i} \succ \hat{c}_{u,j} \mid \Theta) \tag{2.6}$$

This likelihood function aims to optimise the value of Area Under the ROC Curve (AUC) (i.e. maximising the probability that venue $i \in \mathcal{V}_u^+$ is ranked higher than venue $j \in \mathcal{V}_u^-$). To optimise the AUC likelihood function, Rendle et al. (2009) approximated the probability function P using the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, where x is a numerical value, so that the likelihood function is differentiable. Then, the objective function of BPR, $\mathcal{J}(\Theta)$, which aims to learn the latent factors of users and venues, $\Theta = \{P, Q\}$, can be formulated as follows:

$$\mathcal{J}(\Theta) = \underset{\Theta}{argmax} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{V}_u^+} \sum_{j \in \mathcal{V}_u^-} ln(\sigma(\hat{c}_{u,i} - \hat{c}_{u,j})) \; - \lambda_p \sum_{u \in \mathcal{U}} \|p_u\|_F^2 - \lambda_q \sum_{i \in \mathcal{V}} \|q_i\|_F^2 \tag{2.7}$$

In Equation (2.7), regularisation terms are added to avoid overfitting, where λ_p, λ_q are regularisation parameters. Rendle et al. (2009) used matrix factorisation to predict $\hat{c}_{u,i}$, the checkin frequency of user u on venue i based on their historical checkins, obtained by calculating the dot product of the latent factors of the user p_u and the venue q_i (similar to Equation (2.3)). Finally, Stochastic Gradient Descent (SGD) is applied to find the latent factors that give a local maximum of the objective function (Equation (2.7)). More details about the optimisation of BPR is described in Algorithm 2.1. In particular, for each iteration (Algorithm 2.1, Lines: 6-7), given a random feedback tuple of user u who has visited venue i, but not visited venue j, $(u, i, j) \in D = \{(u, i, j) | i \in \mathcal{V}_u^+ \wedge j \in \mathcal{V}_u^-\}$, the parameter $\theta \in \Theta$ is updated based on the gradient of its corresponding parameter $\frac{\partial \mathcal{J}}{\partial \theta}$, while fixing the others, until convergence, as follows:

$$\theta^{(\mathcal{T}+1)} = \theta^{(\mathcal{T})} + \eta^{(\mathcal{T})} \cdot \frac{\partial \mathcal{J}}{\partial \theta}(\theta^{(\mathcal{T})}) \tag{2.8}$$

where $\mathcal{T}$ is the iteration number. The gradients of the latent factor matrices p_u, q_i, q_j are calculated as follows:

$$\frac{\partial \mathcal{J}}{\partial p_u} = \delta(\hat{c}_{u,j} - \hat{c}_{u,i})(Q_i - Q_j) - \lambda_p P_u \tag{2.9}$$

$$\frac{\partial \mathcal{J}}{\partial q_i} = \delta(\hat{c}_{u,j} - \hat{c}_{u,i})P_u - \lambda_q Q_i \tag{2.10}$$

$$\frac{\partial \mathcal{J}}{\partial q_j} = \delta - (\hat{c}_{u,j} - \hat{c}_{u,i})P_u - \lambda_q Q_j \tag{2.11}$$

Algorithm 2.1: An Optimisation Algorithm of BPR

```
1  Input: users $\mathcal{U}$, venues $\mathcal{V}$ and visited venues $\mathcal{V}_u^+$ and unvisited venue $\mathcal{V}_u^-$ for each
     $u \in \mathcal{U}$.
2  Output: $\Theta = \{P \in \mathcal{R}^{m \times d}, Q \in \mathcal{R}^{n \times d}\}$
3  $P \sim U(0,1), Q \sim U(0,1)$
4  repeat
5      for $u \in \mathcal{U}$ do
6          $i \leftarrow$ draw a random visited venue from $\mathcal{V}_u^+$
7          $j \leftarrow$ draw random unvisited venues from $\mathcal{V}_u^-$
8          Compute gradients of $p_u$, $q_i$, $q_j$ as per Equations (2.9 - 2.11)
9          Update the above parameters as per Equation (2.8)
10 until convergence;
```

In this section, we have described the basic Collaborative Filtering-based (CF) approaches for recommendation systems including the *memory-based* and *model-based* CF approaches. The *memory-based* CF approaches predict the rating of a target user on a given venue based on the cosine similarity between the target user and his/her like-minded users who rate similar venues, while the *model-based* approaches (i.e. MF and BPR) aim to maximise (minimise) their objective (loss) function based on the users' observed feedback. In the following section, we describe several types of evaluation methodology as well as metrics commonly used to evaluate the effectiveness of recommendation systems.

2.1.2 Evaluation of Recommendation Systems

The evaluation of recommendation systems plays an important role in the information retrieval development and has been an active research topic (Ricci et al., 2015a; Cañamares and Castells, 2018; Bellogín et al., 2017). In the past decade, various collaborative filtering-based approaches have been proposed in previous literature (Koren et al., 2009; Rendle et al., 2009; He et al., 2017; Ma et al., 2011). Making a decision to select the most appropriate approach for recommendation systems from a large variety of candidate algorithms is a challenging task. In general, such a decision is based on experiments by comparing the performance of a set of candidate algorithms. Furthermore, researchers who propose new recommendation algorithms must also compare the performance of their proposed algorithms with existing approaches. In this section, we discuss how to compare different recommendation algorithms based on comparative studies. In particular, we describe several types of evaluation methodologies commonly used in the recommendation system research communities. We also review various well-known evaluation metrics proposed in previous literature, which are used to select the best performing candidate algorithm for a recommendation system.

2.1.2.1 Evaluation Methodology

There are three types of evaluation methodology for recommendation systems: namely user studies, online and offline evaluations. User studies and online evaluations usually require real users during the evaluation, although they are conducted in different ways. The main difference between these two settings is how the users are recruited. With a user study, a small number of users are asked to interact with the system and asked for feedback before, during or at the end of the study. An online evaluation is usually conducted on a deployed system where a large number of real users are using the system. In contrast, offline evaluations do not require real users and are often easier and less expensive to conduct than those aforementioned two settings. In the following, an overview of these different types of evaluation methodologies is provided.

2.1.2.1.1 User Studies Many recommendation systems require the interactions of users to evaluate their performances. In order to properly evaluate such systems, real users' interactions with the systems must be collected. A user study is conducted by recruiting a small group of users to perform particular tasks and interact with the recommendation systems (Knijnenburg et al., 2012; Sinha et al., 2001). While the users are performing the tasks, their behaviour data with the systems is usually collected and later analysed. Furthermore, the users can be asked to fill questionnaires before, during and after the tasks are completed. A typical example of a user study could be the testing of the influence of a recommendation algorithm on the browsing behaviour of recommendation systems (e.g. (Amatriain et al., 2009; Cosley et al., 2003; Miller et al., 2003; Pu and Chen, 2007)). The users may be asked to check the quality of the ranked list of venues and whether the recommended venues are relevant to their preferences. It is also possible to collect the users' interactions such as how many times a recommendation was clicked as well as their eye movement to track which part of an interface the users was looking at. The advantage of user studies is that we can collect a wealth of information about the users' interactions with the system in a controlled environment. Various experiments can be conducted to evaluate the performance of different recommendation algorithms as well as the effectiveness of the user interface of the recommendation system. However, the main disadvantage of user studies is the active awareness of users about the experiments, which can bias their actions and decisions (Knijnenburg et al., 2012; Sinha et al., 2001). In addition, it is difficult and expensive to recruit a larger group of users. In many cases, the recruited users may not be representative of the intended audiences of the recommendation system. Therefore, the results from such user evaluations cannot be fully trusted (Ricci et al., 2015a). In this thesis, we will not evaluate the effectiveness of our proposed approaches using user studies due to its aforementioned disadvantages.

2.1.2.1.2 Online Evaluations Online evaluations bear some similarities with user studies, except that the users are typically real users in a fully deployed or commercial system. Therefore, there is less bias from the recruitment process, unlike the user studies, because the users directly use the system in the natural course of their activities. Hence, these users can be representative of the general population. With online evaluation, we can evaluate the comparable effectiveness of recommendation systems by comparing various recommendation algorithms (Kohavi et al., 2009). In particular, users can be sampled randomly and each recommendation algorithm can be evaluated with each sample of users. This method is referred to as A/B testing, and it measures the direct impact of the recommendation algorithms on the end user. User engagement metrics such as page views, click-through rate (Garcin et al., 2014) and the economic benefit of the system (Shani et al., 2005) can be used to evaluate the effectiveness of such recommendation algorithms. The basic idea in A/B testing is to compare two algorithms as follows:

1. Randomly divide the users into two groups (A and B)

2. Deploy one recommendation algorithm for group A and another algorithm for group B for a period of time, while keeping all other conditions across the two groups (e.g. user-interface) as similar as possible.

3. At the end of the period of time, compare the performance of the two algorithms using particular metrics.

However, the main disadvantage of A/B testing is the risk involved in performing evaluations on the real system, as a tested under-performing recommendation algorithms may affect negatively the experience and satisfaction of real users. Another disadvantage of this type of evaluation is that we can only get a comparable evaluation between two systems rather than an absolute evaluation (Kharitonov, 2016).

2.1.2.1.3 Offline Evaluations In the offline evaluations, pre-collected historical data such as ratings and checkins data collected from real-world commercial LBSNs are used to evaluate the effectiveness of recommendation systems. Such historical data may also be associated with additional information such as the textual content of comments associated with the ratings left by the users and temporal information (e.g. timestamp) about when each user has rated the item. Using these data, we can simulate the users' behaviour when interacting with a recommendation system. An assumption of the offline evaluations is that the user behaviour when the data was collected and when the recommendation system is deployed are similar, so that we can make reliable decisions based on the simulation (Ricci et al., 2015a). Offline evaluations have been the most widely accepted techniques for recommender system evaluation because the statistical robustness and explainable quantifications provided by the offline evaluations (Aggarwal et al., 2016; Vargas, 2015). The main advantage of the use of historical data is that they do not require real user engagement and thus allow us to compare various recommendation algorithms at a low cost. Once the data has been collected, it can be used as a standard benchmark to compare various recommendation systems across various settings. Furthermore, historical data from various platforms that exhibit difference user's behaviour (e.g. rating dataset from Yelp and checkin dataset from Foursquare) can be used to investigate the generalisation of the recommendation systems. However, unlike the online evaluations, the main disadvantage of offline evaluation is that it cannot not guarantee the actual preferences of the users in the future. For instance, the collected data is static and the current recommendations may not reflect the most appropriate recommendations for the future as the actual preferences of the users evolve. Ideally, the goal of the offline evaluations is to filter out under-performing recommendation algorithms, leaving a relatively small set of candidate algorithms to be tested by the more costly user studies or online experiments (Ricci et al., 2015a). Moreover, we can use offline evaluation to turn the parameters of the algorithms and then further conduct either user studies or online evaluation to evaluate the algorithm with the best tuned parameters. In this thesis, we evaluate the effectiveness of our proposed approaches using the offline evaluations due to its simplicities. In the next section, we describe various evaluation metrics widely used to evaluate the effectiveness of recommendation systems.

2.1.2.2 Evaluation Metrics

A large number of evaluation metrics have been proposed to evaluate the effectiveness of venue recommendation systems (Järvelin and Kekäläinen, 2002; Deshpande and Karypis, 2004; Hurley and Zhang, 2011; Lu et al., 2012). The effectiveness of venue recommendation systems can be evaluated by measuring the accuracy of predicting rating values (e.g. with Root Mean Square Error and Mean Absolute Error) or by measuring the accuracy of the recommended ranked lists of venues (e.g. Precision, Recall and Normalised Discounted Cumulative Gain). Apart from the accuracy metrics, there are several aspects to evaluate the recommendation systems such as novelty, serendipity as well as diversity (Hurley and Zhang, 2011; Lu et al., 2012; Vargas, 2015). However, these metrics are not appropriate to evaluate the effectiveness of our proposed approaches. In this thesis, we focus on enhancing the recommendation accuracy of recommendation systems. Indeed, Vargas (2015) argued that the recommendation systems that are optimised for the accuracy metrics do not necessarily improve the novelty and diversity of the recommendations. Since all of our proposed approaches are optimised for the accuracy metrics, using the novelty, serendipity and diversity metrics to evaluate the effectiveness of our proposed approaches in comparison with baselines are not appropriate because our proposed approaches as well as the baselines are optimised for the accuracy metrics.

The accuracy metrics can be categorised into two categories: namely rating prediction- and ranking-based metrics. Rating prediction-based metrics are used to evaluate the prediction accuracy of users' ratings on particular venues by recommendation systems. Let $r_{u,i}$ be the value of the rating of user u on venue i, which is used in the test set, $(u, i) \in E$, (i.e. ground-truth data) and $\hat{r}_{u,i}$ be the predicted rating by a specific training algorithm. Two rating prediction-based metrics widely used in previous literature (Ma et al., 2011; Guo et al., 2015b; Hu et al., 2014; Koren et al., 2009) are Mean Absolute Error (MAE) and Root Mean

Square Error (RMSE) (for both metrics, lower is better), which are calculated as follows:

$$MAE = \frac{\sum_{(u,i)\in E} \mid \hat{r}_{u,i} - r_{u,i} \mid}{\mid E \mid} \tag{2.12}$$

$$RMSE = \sqrt{\sum_{(u,i)\in E} \frac{(\hat{r}_{u,i} - r_{u,i})^2}{\mid E \mid}} \tag{2.13}$$

Ideally, MAE measures the average magnitude of the errors in a set of predictions, while RMSE is the square root of the average of squared differences between the predicted ratings and the actual ratings (i.e. errors). Taking the square root of the average squared errors has some interesting implications for RMSE over the MAE metric. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. In particular, RMSE tends to disproportionately penalise large errors because of the squared term within the summation, while MAE does not disproportionately penalise larger errors. In particular, as RMSE sums up the squared errors, it is more significantly affected by large error values (i.e. a few severely inaccurate predicted ratings can markedly increase the error value reported by the RMSE measure). In a scenario where the robustness of prediction across various ratings is important, RMSE is more appropriate than MAE (i.e. RMSE is more useful than MAE when large errors are particularly undesirable.). However, the main disadvantage of RMSE is that it cannot reflect average errors, which can lead to misleading results (Willmott and Matsuura, 2005; Chai and Draxler, 2014).

In contrast to the rating prediction-based metrics, which focus on the errors between the predicted ratings and the ground-truth ratings, the ranking-based metrics focus on the accuracy of the ranks of the top-K venues. Indeed, the evaluation of the quality of rankings is generally measured by Information Retrieval metrics such as Precision, Recall, Mean Average Precision (MAP), Normalized Discount Cumulative Gain (NDCG), Mean Reciprocal Rank (MRR) and Hit Ratio (HR).

Precision is the fraction of the recommended venues that are relevant to the user's preference, while recall is the fraction of relevant venues that have been recommended over the total amount of relevant venues. In particular, precision measures how good our recommended venues are and recall measures how many of the good venues are recommended, in comparison to how many there are. Both precision and recall are binary metrics (i.e. considering whether a given item is relevant or not) and the higher is the better. As mentioned in Section 1, users on the LBSNs only focus on venues ranked at the top positions (Yuan et al., 2016; Ying et al., 2016; Li et al., 2015), we calculate precision and recall for the first K venues instead of all venues for each user, denoted as Pre@K and Rec@K, which are calculated as follows:

$$Pre@K = \frac{\sum_{i \in R_u} rel_u(i)}{|R_u|} \tag{2.14}$$

$$Rec@K = \frac{\sum_{i \in R_u} rel_u(i)}{\sum_{i \in \mathcal{V} rel_u(i)}} \tag{2.15}$$

Let R_u be the set of top N recommendations for user u where R_u is the set of top-K recommendations of user u, $rel_u(i)$ returns 1 if venue i is relevant to user u, otherwise 0. Both precision and recall do not consider the actual rank of the correct recommendations among those recommended. MAP is the mean of the Average Precision (AP) values averaged over all users. The AP for a user is the average of all precision values calculated after each item is recommended (Voorhees and Harman, 2003). Note that MAP is a top-heavy measure, i.e. venues ranked correctly near the top of the ranking contribute more to the MAP performance than venues ranked near the bottom. MAP is calculated as follows:

$$MAP = \sum_{u \in \mathcal{U}} \frac{\sum_{k=1}^{K} Pre(R(u), k) \cdot Rec(R(u), k)}{| \mathcal{U} |} \tag{2.16}$$

where k is a rank of the recommended item $R(u)$ for user u, $Pre(R(u), k)$ is the precision at cut-off k and $Rec(R(u), k)$ is the change in recall between rank $k - 1$ and k.

Another ranking-based metric widely used in recommendation systems is the Mean Reciprocal Rank metric (MRR) proposed by Voorhees et al. (1999) that measures the rank of the first relevant item in the item list averaged over all users, $\mathcal{U}$, which is calculated as follows:

$$MRR = \frac{1}{| \mathcal{U} |} \sum_{i=1}^{|U|} \frac{1}{rank_i} \tag{2.17}$$

where $rank_i$ is the rank position of the first relevant item for i-th user and U is a set of all users. However, MRR shows some strange behaviour in some scenarios as reported by Fuhr (2018). For example, given three users, system A recommends the first relevant venue for these three users at ranks 1, 2, and 4, respectively, while system B recommends the relevant venues for each user at rank 2. Therefore, the average of the first relevant venues of system A is 2.33 ($\frac{1}{3}(1+2+3)$), which is worse than that system B does (i.e. rank 2 on the average). However, the MRR score of system A is 0.58 ($\frac{1}{3}(\frac{1}{1} + \frac{1}{2} + \frac{1}{4})$)), while the system B's MRR score is 0.5 ($\frac{1}{3}(\frac{1}{2} + \frac{1}{2} + \frac{1}{2})$)), hence system A outperforms system B on MRR and is outperformed by system B on the average rank. A simplification of MRR is the Hit Ratio (HR), a recall-based metric proposed by Deshpande and Karypis (2004), in which the rank reciprocal weighting is not used, and the value of HR for each user is either 1 or 0. The HR metric has been commonly used in top-N evaluation for recommendation systems (He et al., 2017; Xiang et al., 2010; Lee et al., 2011) when the ground-truth data are extracted from the implicit feedback. In particular, HR is simply the fraction of users for which the relevant item is included in the recommendation list of length K (i.e. $HR@K = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} H_u$, where H_u returns 1 if at least one relevant venue appears in the recommendation of user u, R_u, otherwise 0). Ideally, if a ground-truth item appears in the recommended list, it is deemed to be a hit.

The main disadvantage of MAP and MRR for the evaluation of recommendation systems is that they are built upon a binary relevance grade (i.e. each recommended item is considered either relevant or non-relevant). Hence, these metrics are not suitable to evaluate the effectiveness of recommendation systems when multiple relevance grades are available in the ground-truth data. For example, venues might be considered in terms of 1-5 rating scales from explicit feedback or in terms of binary scales from implicit feedback. To evaluate recommendation systems with multiple relevance grades within ground-truth data, we can use a Discounted Cumulative Gain (DCG) metric as proposed by Järvelin and Kekäläinen (2002) to measure the performance of recommendation systems based on the graded relevant of the recommended venues, which is calcuated as follows:

$$DCG = \sum_{i=1}^{K} \frac{2^{rel_i}}{\log_2 i + 1} \tag{2.18}$$

where rel_i represents the ground-truth relevance score of the candidate venue at the position i, where $rel_i = 1$ if the candidate venue is relevant, otherwise $rel_i = 0$. Then, the normalised NDCG ($NDCG$) is calculated as follows:

$$NDCG = \frac{DCG}{IDCG} \tag{2.19}$$

where the ideal DCG ($IDCG$) is a score that would have been achieved by the perfect ranking according to the ground-truth data.

In this section, we have described a number of evaluation metrics that have been widely used in the literature to evaluate the effectiveness of recommendation systems (Järvelin and Kekäläinen, 2002; Deshpande and Karypis, 2004; Hurley and Zhang, 2011; Lu et al., 2012). Later in Chapters 4-7, we use these metrics to evaluate the effectiveness of our proposed approaches in comparison with various existing venue recommendation baselines. In the next section, we describe several basic architectures of Deep Neural Networks such as Multi-Layer Perceptron, Convolutional Neural Networks and Recurrent Neural Networks.

2.2 Basic Architecture of Deep Neural Networks

Deep Neural Network (DNN) models are popular machine learning techniques that simulate the mechanism of a human nervous system, which are referred to as neurons. Each neuron is connected to each other in order to transfer information and learn complex correlation of given data. Previous literature have been shown to be promising and impressive successes of DNN models have been observed in domains such as speech recognition, computer vision and natural language processing (e.g. (He et al., 2016a; Zhang et al., 2014a; Kim, 2014)) as well as recommendation systems (e.g. (Yu et al., 2016; He et al., 2017; He and Chua, 2017; Cheng et al., 2016; Liu et al., 2016b). In this section, we provide an overview of the basic architecture of various DNN models including Multi-Layer Perceptron (Section 2.2.1), Convolutional Neural Networks (Section 2.2.2) and Recurrent Neural Networks (Section 2.2.3).

2.2.1 Multi-Layer Perceptron (MLP)

In this section, we describe single-layer and multi-layer neural networks. In a single-layer network, a set of inputs is directly mapped to an output by using a linear function (Aggarwal, 2018, p.5). This simple architecture of neural network is also referred as a perceptron, which is defined as follows:

$$\hat{y} = a(Wx + b) \tag{2.20}$$

where $a()$ is called the activation function. In a Multi-Layer Perceptron (MLP), the perceptrons are arranged in multiple layers and they are connected to each other (i.e. the output of perceptron in the first layer is the input of perceptron in the next layer), which is defined as follows:

$$\hat{y}_{MLP} = a_L(W_L(...a_1(W_1x + b_1)) + b_L) \tag{2.21}$$

where L is a number of layers of MLP. Note that different types of activation functions such as the sigmoid and hyperbolic tangents may be used in different layers. The choice of activation function plays an crucial role for neural network design. For example, if it is a regression task where the predicted values are real numbers, a linear activation function ($f(x) = ax$, where a is a numerical parameter) is preferred (Aggarwal, 2018, p. 11-14). Unlike the linear activation, nonlinear activation functions can map the outputs from an arbitrary range to bounded outputs. For example, the sigmoid activation ($f(x) = \frac{1}{1+e^{-x}}$) outputs a value between 0 and 1, which is suitable for computing probabilities (Aggarwal, 2018, p. 11-14). The hyperbolic tangent (tanh) function is similar to the sigmoid function, except that it outputs a value between -1 and 1. The tanh function is preferable to the sigmoid function when the outputs of the computations are desired to be both positive and negative.

In a single-layer neural network, the training process is relatively straightforward because we can apply the Stochastic Gradient Descent (SGD) technique to update the parameters of the network, $W_1, b_1, ..., W_L, b_l$. However, for multi-layer networks, the loss function is a complicated composition function of the parameters in earlier layers. Therefore, to compute the gradients of MLP, a back-propagation algorithm that exploits the chain rule of differential calculus is needed. In particular, the back-propagation consists of two main phases, namely the forward and backward propagation phases. Given training instances, the forward phase computes the predicted values and the local derivatives with respect to the weights of all layers. Then, in the backward propagation phase, we accumulate the products of these local gradients over all layers in a backward direction (i.e. from the output layer to the first layer) in order to update the weights of each layer with respect to the local gradients. Later, in Chapter 5, we will demonstrate how to exploit MLP to effectively capture the complex structure of the user-venue interactions in order to enhance the quality of venue recommendations.

The main advantage of MLP over a single-layer neural network is that MLP can effectively capture the dependencies between each dimension of input vector. In addition, with the variety of activation functions of each layer in Equation (2.21) (i.e. $a_1...a_l$) MLP can map the outputs of previous layers from an arbitrary range to bounded outputs, which are suitable for computing probabilities (Skansi, 2018). Note that MLP supports both regression and classification problems by using appropriate loss function and activation function. For example, sigmoid function and binary cross entropy loss function are appropriate for binary classification problem, while linear function and mean square error loss function are appropriate for regression problem. However, MLP is not suitable to model the dependencies between each dimension of 2D and 3D matrix inputs. In the next section, we describe the Convolutional Neural Networks that are suitable for such matrix inputs.

2.2.2 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) are biologically inspired networks that are used in computer vision for image classification and object detection (Aggarwal, 2018, p. 40-44). CNN are originally designed to work with grid-structured inputs, which aim to capture spatial relationship in local regions of the grid. The most obvious example of grid-structured data is a 2-dimensional image. This type of data also exhibits spatial relationship between the regions because adjacent spatial locations in an image often have similar colour values in their individual pixels. Apart from image processing tasks, CNN have also been applied to text classification problems in the literature (e.g. (Kim, 2014; Yang et al., 2016; Zhang

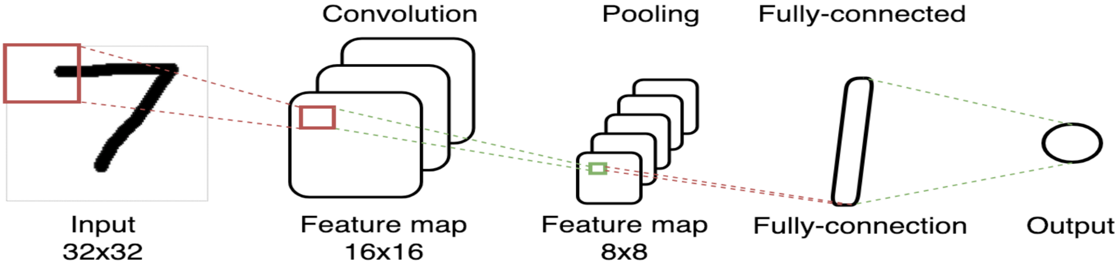


Figure 2.1: The architecture of Recurrent Neural Networks. This figure is inspired from Teow (2017)

et al., 2016)). In recommendation systems, various approaches (Tang et al., 2015; He and McAuley, 2016; Zheng et al., 2017)) have been proposed to exploit CNN to leverage textual information and images to enhance the accuracy of rating prediction and the quality of recommendations. For simplicity, we describe the basic architecture of CNN using the digit handwriting image classification problem as an example scenario where the input of the CNN is 32x32 matrix and the task is to predict a digit of a given image (i.e. a multi-class classification problem). The basic architecture of a CNN is illustrated in Figure 2.1. A CNN is relatively similar to MLP, except that the operations in its layers are spatially organised with sparse connections between layers. The architecture of a CNN consists of four main types of layers including convolution, pooling, fully-connected and prediction layers. The convolution layer maps each region of the given image into the smaller matrix (i.e. from 32x32 to 16x16 feature map) using a kernel/filter. Then, each feature map is passed to the pooling layer, which essentially downsample a number of neurons in the feature map based on the pooling operation (e.g. max or min pooling). Next, the pooled feature maps are passed to the fully-connection and output layer, sequentially. The prediction layer is often fully connected and maps in an application-specific way. For example, with the classification problem, the softmax activation can be deployed.

2.2.3 Recurrent Neural Networks (RNN)

All of the aforementioned neural architectures such as the Multi-Layer Perceptron (MLP) and the Convolutional Neural Networks (CNN) were originally designed for multi-dimensional data in which the attributes are largely independent of one another. However, these neural architectures are not suitable for certain data types such as time-series and text data, which exhibit sequential patterns. For example, in time-series data (e.g. currency exchange rates and house prices), the values on successive timestamps are closely related to one another. If one uses the values of these timestamps as independent features, then the information about

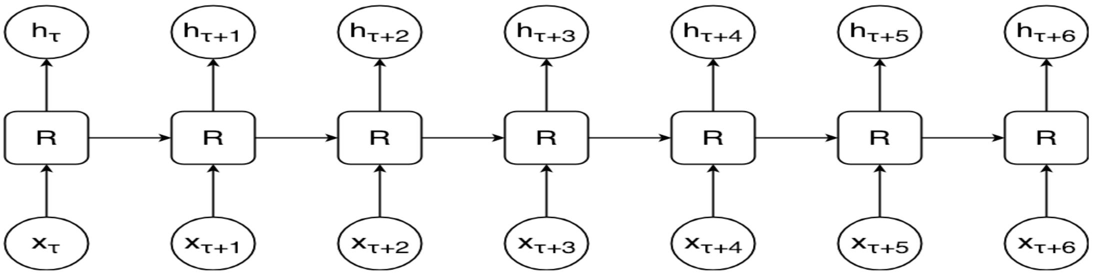


Figure 2.2: The architecture of Recurrent Neural Networks

the relationships among the values (i.e. sequential properties) might be missing. Recurrent Neural Network (RNN) models are neural networks that take information about the previous neural network unit into account. RNN models are suitable and have been incredibly successful when applied to problems where the input data on which the predictions are to be made is in the form of a sequence (e.g. a sequence of words for question answering or a sequence of checkins for next venue prediction) (Yu et al., 2016; Beutel et al., 2018; Liu et al., 2016c; Donkers et al., 2017; Smirnova and Vasile, 2017). In the following, we describe the basic architecture of three popular recurrent neural networks: namely traditional RNN, Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU).

2.2.3.1 Architecture of Traditional Recurrent Neural Networks

The basic architecture of a traditional recurrent neural network is illustrated in Figure 2.2. The illustration shows a recurrent unit R (a rectangle) and the arrows between each unit R to indicate sequential information that is passed to the successive units. Circle x_τ and h_τ indicate the input and output of the recurrent unit at time step τ, respectively. In particular, given an input vector $x_\tau \in \mathbb{R}^d$ at each time step τ, where d is the dimensions of the input vector, the output of the recurrent unit, h_τ is calculated as follows:

$$h_\tau = \sigma(Wx_\tau + Rh_{\tau-1}) \tag{2.22}$$

where $h_{\tau-1}$ is the output of previous recurrent unit, $\sigma(x)$ is the sigmoid function and W is a transition matrix and R is a recurrent connection weight matrix that captures sequential signals between every two adjacent hidden states h_τ and $h_{\tau-1}$.

The main advantage of RNN models is that they take the inputs from the previous units into account, which might capture useful sequential properties. However, traditional RNN models usually suffer from the vanishing gradient problem when the models are trained on long sequences of inputs (Hochreiter and Schmidhuber, 1997; Chung et al., 2014). In

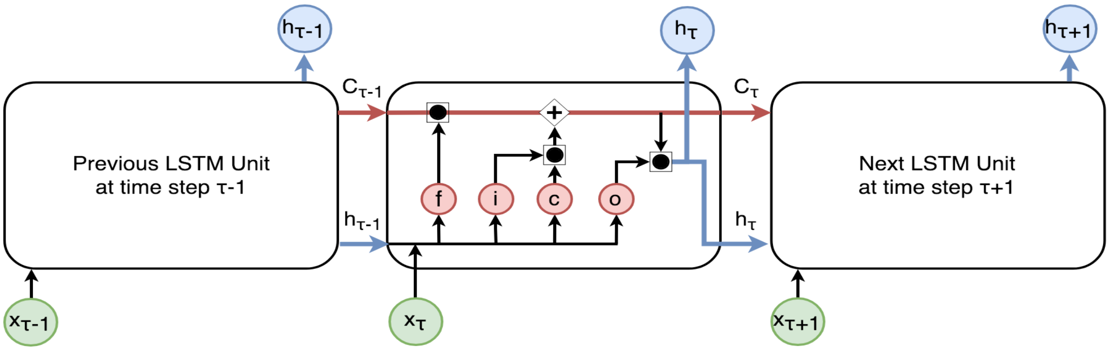


Figure 2.3: The architecture of Long Short-Term Memory

particular, as the gap between the recurrent units[1] increase (i.e. recurrent unit τ and $\tau + 6$ in Figure 2.2), RNN models become unable to learn to connect the information. Note that the vanishing gradient problem is a common problem in neural network parameter updates where successive multiplications of the matrix W of successive recurrent units is not stable (see Equation (2.22)). This can cause either the gradient to disappear during the back-propagation or gradient values are fluctuated.

2.2.3.2 Long Short-Term Memory (LSTM)

As discussed in Section 2.2.3, traditional RNN models usually suffer from the vanishing gradient problem when dealing with long sequences of inputs (Hochreiter and Schmidhuber, 1997; Chung et al., 2014). To alleviate the vanishing gradient problem of the traditional RNN models, Hochreiter and Schmidhuber (1997) introduced Long Short Term Memory networks (LSTM), which are an extension of the traditional RNN models, which uses gating mechanisms to learn long- and short-term dependencies in long sequences of inputs. Figure 2.3 illustrates the architecture of LSTM. The green and blue circles denote the input and output of each LSTM unit, respectively. An important component of LSTM is the cell state, C_τ, the top red horizontal line running through the LSTM units. The cell state can be seen as a long-term memory that retains at least a part of the information from previous LSTM units. The gating mechanism of each LSTM unit consists of a forget gate (f), an input gate (i), a candidate cell state (c) and an output gate (o), denoted as red circles inside the LSTM unit. These gates carefully regulate the flow of information to the cell state. The first step of the gating mechanism is to decide what information can be ignored or throw away from the cell

[1] The gap between two recurrent units is a number of recurrent units located between those two recurrent units.

state. This decision is made by the forget gate as follows :

$$f_\tau = \sigma(W_f x^\tau + R_f h_{\tau-1} + b_f) \tag{2.23}$$

where W, R, b are the transition matrix, the recurrent matrix and the corresponding bias parameters and $\sigma()$ is the sigmoid function. The inputs of the forget gate are input x_τ at current unit at time step τ and hidden state of previous unit h_τ. The output of the forget gate is between 0 and 1, where 1 represents "completely keep the current input and the hidden state from previous unit", while 0 represents "completely ignore them". The next step is to decide what new information will be memorised in the cell state. This step requires the input gate and candidate cell state, which are defined as follows:

$$i_\tau = \sigma(W_i x^\tau + R_i h_{\tau-1} + b_i) \tag{2.24}$$

$$\widetilde{c}_\tau = tanh(W_c x^\tau + R_c h_{\tau-1} + b_c) \tag{2.25}$$

where $tanh()$ is a hyperbolic tangent function. Then, the previous cell state, $C_{\tau-1}$, can be updated into the new cell state, C_τ, based on the outputs of the forget and input gates as well as the candidate cell state, which is defined as follows:

$$C_\tau = f_\tau C_{\tau-1} + i_\tau \widetilde{c}_\tau \tag{2.26}$$

Ideally, the values of new cell state C_τ are determined by the candidate cell state $\widetilde{c}_\tau$ and the cell state of the previous LSTM unit $C_{\tau-1}$, while the influences of $\widetilde{c}_\tau$ and $C_{\tau-1}$ are controlled by the input gate i_τ and forget gate f_τ, respectively. Finally, the output of the current LSTM unit at time step τ, h_τ is calculated based on the updated cell state, C_τ and the output of the output gate, which are defined as follows:

$$o_\tau = \sigma(W_o x^\tau + R_o h_{\tau-1} + b_o) \tag{2.27}$$

$$h_\tau = o_\tau \cdot tanh(c_\tau) \tag{2.28}$$

Using the gating mechanisms that include the forget gate, input gate, candidate cell state and output gate, LSTM can effectively control the influences of the hidden state of previous LSTM unit, and hence can alleviate the vanishing gradient problem when dealing with long sequences of inputs (Hochreiter and Schmidhuber, 1997). Recently, LSTM has been successfully applied in many areas where the inputs are long such as text classification (Zhou et al., 2015; Lee and Dernoncourt, 2016; Zhou et al., 2016). In the following section, we describe an extension of LSTM that can alleviate the vanishing gradient problem, while has

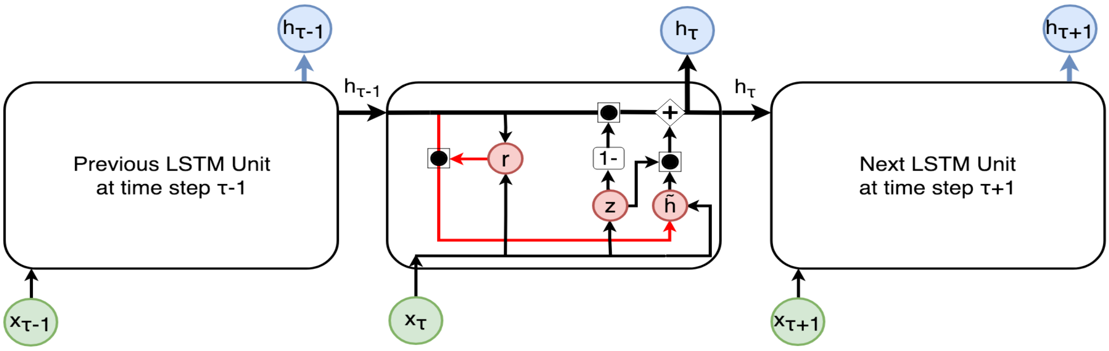


Figure 2.4: The architecture of GRU units

less parameters than the LSTM.

2.2.3.3 Gated Recurrent Units (GRU)

The Gated Recurrent Unit (GRU) proposed by Chung et al. (2014) can be viewed as a simplification of the LSTM, which does not require the cell states. Unlike LSTMs, which directly use the forget and output gates to control the amount of information changed in the hidden state, the GRU unit uses a single reset gate to achieve the same goal. However, the basic idea of the GRU is relatively similar to that of LSTM, in terms of how it partially resets the hidden states. Although GRU is a closely related simplification of LSTM, it should not be seen as a special case of LSTM. Comparisons of LSTM and GRU have been conducted in (Chung et al., 2014; Jozefowicz et al., 2015), which both showed that these two models are similar in performance, and the relative performance seems to depend on the task at hand. Indeed, the main advantage of GRU over LSTM is its memory and training efficiency (i.e. less parameters) as well as its ease of implementation. In addition, GRU is more effective than LSTM when the training data is not largely available because of a smaller parameter footprint (Chung et al., 2014; Greff et al., 2017), although LSTM would be preferable with an increased amount of training data. Figure 2.4 illustrates the architecture of GRU units. Similar to LSTM, GRU consists of gating mechanisms that control the influence of the hidden state of the previous unit $h_{\tau-1}$ in the current unit at time step τ. Indeed, GRU can learn to ignore the previous units if necessary. The gating mechanism of GRU unit is defined as follows:

$$z_\tau = \sigma(W_z x^\tau + R_z h_{\tau-1} + b_z) \tag{2.29}$$

$$r_\tau = \sigma(W_r x^\tau + R_r h_{\tau-1} + b_r) \tag{2.30}$$

$$\widetilde{h}_\tau = \tanh(W x^\tau + R(r_\tau \odot h_{\tau-1})) \tag{2.31}$$

$$h_\tau = (1 - z_\tau)h_{\tau-1} + z_\tau \widetilde{h}_\tau \tag{2.32}$$

where z_τ, r_τ are update and reset gates, respectively, and $\widetilde{h}_\tau$ is a candidate hidden state. respectively. R is a recurrent connection weight matrix that captures sequential signals between every two adjacent hidden states h_τ and $h_{\tau-1}$, using $\odot$, which denotes the element-wise product. Later, in Chapter 3, we will review a number of existing approaches proposed in the literature (e.g. Zhu et al. (2017); Smirnova and Vasile (2017); Beutel et al. (2018)) that exploit the GRU units to model the users' preferences from their sequences of observed feedback in order to improve the quality of context-aware venue recommendations.

2.3 Conclusions

In this chapter, we have provided a summary of the key concepts of recommendation systems and described some basic architectures of deep neural networks that this thesis builds upon. In particular, in Section 2.1, we described the general background of recommendation systems including the goals and formulations of recommendation systems, while Section 2.1.1 detailed collaborative filtering algorithms. In Section 2.1.2, we discussed evaluation in recommendation systems, including the evaluation methodologies and metrics used for evaluating the effectiveness of recommendation systems that we use in later chapters. Finally, Section 2.2 described the basic architectures of deep neural networks such as Multi-Layer Perceptron (MLP) and Recurrent Neural Networks (RNN), which will be used in later technical chapters. In the next chapter, we discuss related work that exploits both collaborative filtering techniques and deep neural networks for venue recommendation system. We also identify the knowledge gap of these related work that this thesis addresses.

Chapter 3

Related Work

3.1 Introduction

In the previous chapter, we provided an overview of recommendation systems including basic collaborative filtering approaches and the used evaluation methodologies as well as basic architectures of deep neural networks that we build on in this thesis. In this chapter, we review related works on venue recommendation systems that have been previously proposed in the literature. We first review several Matrix Factorisation-based approaches that leverage additional information such as the users' friendships, the geographical information of venues and the textual content of comments to enhance the user-venue rating prediction accuracy. Then, we describe the state-of-the-art Matrix Factorisation framework that exploits Deep Neural Networks (DNN) to model the user-venue interactions for the user-venue rating prediction task. On the other hand, as mentioned in Section 1.1, users in LBSNs only focus on the top-K ranked list of venues for obtaining recommendations, and hence the MF-based approaches that aim to minimise Root Mean Square Errors between the predicted ratings and the observed ratings may not provide an effective top-K ranked list of venues. We discuss the existing negative sampling approaches and ranking function that aim to enhance the effectiveness of Bayesian Personalised Ranking (BPR) for effective top-K venue recommendations. Next, we provide details of existing RNN-based factorisation approaches that exploit RNN models to capture the users' short-term (*dynamic*) preferences for the sequential order of checkins. Finally, as mentioned in Section 1.1, the users' current context (e.g. their location and time of the day) play an important role in Context-Aware Venue Recommendation (CAVR), as it can influence the users' decision to visit venues. We discuss extensions of the traditional GRU architectures proposed in previous works that enable the GRU units to incorporate contextual information associated with the sequence of checkins for CAVR. Note that, although some of these factorisation-based approaches, negative sampling approaches and recurrent architectures were not originally proposed for CAVR, they are sufficiently flexible to be applied to the CAVR task. For consistency, in this section, we explain these approaches in the domain of CAVR. Apart from the description of these approaches, in this chapter, we also discuss their advantages and disadvantages and thereby identify a series of limitations of these approaches that we aim to address in the later chapters. The outline of this chapter is detailed as follows:

- Section 3.2 formalises the problem statement of rating prediction, top-K venue recommendation and context-aware venue recommendation. This section also introduces notations used in this chapter and the remaining chapters of this thesis.
- Section 3.3 reviews the MF-based approaches previously proposed in the literature (Hu et al., 2014; Jin et al., 2016; Ma et al., 2011) that leverage additional information, such as social and textual information, to enhance the prediction accuracy of a traditional MF model. These MF-based approaches can be categorised into two categories: namely regularisation techniques (Section 3.3.1) and factorisation approaches (Section 3.3.2). We also identify the limitations of these MF-based approaches.
- Section 3.4 discusses the key limitation of traditional MF-based approaches, identified by previous work (He et al., 2017), which rely on the dot product of latent factors to generate the venue recommendations. Then, we review the Neural Matrix Factorisation framework (NeuMF) proposed by He et al. (2017) that aims to address this key limitation of the traditional MF-based approach. In particular, the NeuMF framework consists of two models: namely the Generalised Matrix Factorisation (GMF) (Section 3.4.3) and the Multi-level Perceptron (MLP) (Section 3.4.4) models that both exploit DNN to capture the complex structure of user-venue interactions. Thereafter, we identify the limitations of the NeuMF framework in Section 3.4.5
- Section 3.5 reviews existing negative sampling approaches and pairwise ranking functions for BPR proposed in previous literature (Wang et al., 2016; Yuan et al., 2016; Zhao et al., 2014). Their proposed negative sampling approaches leverage additional information such as the geographical information of venues and social links between the users of LBSNs to enhance the effectiveness of BPR for venue recommendation. Section 3.5.3 discusses the limitations of these approaches.
- Section 3.6 describes the existing RNN-based factorisation approaches (Yu et al., 2016; Zhang et al., 2014b; Tang et al., 2017) that combine RNN models and the traditional MF-based approach to capture the sequential properties of users from their

Table 3.1: Summary of venue recommendation tasks

Task	Input	Output
Rating prediction	user u and venue i	predicted rating $\hat{r}_{u,i}$
Top-K venue recommendation	user u	ranked list of venues
Context-aware venue recommendation	user u and context such as his/her current geographical location, time of the day and historical sequence of checkins $s_{u,t}$	ranked list of venues

checkin sequences. Next, in Section 3.6.2, we discuss the limitation of traditional GRU architecture identified by previous works (Zhu et al., 2017; Beutel et al., 2018; Smirnova and Vasile, 2017) and their proposed extensions of the GRU architecture that incorporate the contextual information associated with sequences of checkins for CAVR. Section 3.6.3 summarises a number of elicited limitations of these RNN-based factorisation approaches and the extensions of GRU architectures.

- Section 3.7 summarises all the elicited limitations of the approaches described in this chapter and discuss how we address these limitations in the later chapters of this thesis.
- Section 3.8 provides concluding remarks for this chapter.

3.2 Problem Statement and Notations

In this section, we describe the problem statement of several venue recommendation tasks as well as introduce the notations used in this thesis. As mentioned in Section 2.1.1.1, venue recommendation can be categorised into two tasks: namely user-venue rating prediction and top-K venue recommendation. Table 3.1 summarises these tasks in details. The task of rating prediction is to predict the rating the user might give to a given venue. Let $\mathcal{U}$ and $\mathcal{V}$ denote the set of all users and venues, respectively, and $R \in \mathbb{R}^{m \times n}$ be the explicit user-venue rating matrix, where $m = |\mathcal{U}|$ and $n = |\mathcal{V}|$ are the number of users and venues, respectively. $r_{u,i} \in R$ indicates the explicit rating feedback by user $u \in \mathcal{U}$ on venue $i \in \mathcal{V}$, typically $0 < r \leq 5$.

In contrast to the rating prediction task, the top-K venue recommendation task aims to generate a ranked list of venues that a user might prefer to visit based on his/her historical implicit feedback (e.g. previously visited venues from checkin data). As described in Table 3.1, the top-K venue recommendations and context-aware venue recommendations are similar ex-

Table 3.2: List of notations used in this thesis.

Notation	Meaning
R	user-venue rating matrix
C	users' checkin matrix
F	users' social link matrix
$\mathcal{U}$	set of users
$\mathcal{V}$	set of venues
$\mathcal{V}_u^+$	list of venues that user u has visited
$\mathcal{S}_u$	sequence of checkins of user u
$r_{u,i}$	rating of user u on venue i
$c_{u,i,t}$	checking of user u on venue i at time t
$s_{u,t}$	sequence of checkins of user u up to time t
lat_j, lng_i	latitude and longitude of venue i
P	the latent factors of users
p_u, ϕu_u	the latent factors of user u projected from P
Q	the latent factors of venues
q_i ϕv_i	the latent factors of venue i projected from Q
ϕt	the latent factors of time
τ	particular time step of the recurrent unit
h	hidden state
Δt	the time interval between two checkins
Δg	the geographical distance between two checkins
σ	the sigmoid activation
$tanh$	the tangent activation
$\odot$	the dot product operation
$\otimes$	the element-wise product operation
$\oplus$	the concatenation operation

cept that the context-aware venue recommendation task takes into account the users' context, while the top-K venue recommendation task does not. Let $C \in \mathbb{R}^{m \times n}$ be the implicit users' checkin matrix, where each entry $c_{u,i,t}$ denotes a user $u \in \mathcal{U}$ who has checked-in into venue $i \in \mathcal{V}$ at timestamp t. Note that $c_{u,i,t} = 0$ means that user u has not made a checkin at venue i at time t. Social links are represented as a matrix $F \in \mathbb{R}^{m \times m}$ where F_u is the set of user u's friends. Let $\mathcal{V}_u^+$ denote the list of venues that the user u has previously visited, sorted by time and let $\mathcal{S}_u$ denote the set of sequence of checkins (e.g. $\mathcal{S}_u = \{[c_1], [c_1, c_2], [c_1, c_2, c_3]\}$). $s_{u,t} = \left\{c = (u, i, \grave{t}) \in \mathcal{C} \mid \grave{t} < t\right\} \subset \mathcal{S}_u$ denotes the sequence of checkins of user u up to time t. We use $s_{u,t}^{\tau}$ to denote the τ-th checkin in the sequence. t^{τ} denotes the timestamp of τ-th checkin. lat_i, lng_i denote the geographical position, in terms of latitude and longitude of checkin/venue i.

3.3 Extensions of Matrix Factorisation Models

Matrix Factorisation (MF) is a collaborative filtering-based approach widely used to predict ratings that users give to venues (Hu et al., 2014; Cheng et al., 2012; Yuan et al., 2016). As described in Section 2.1.1.1.2, traditional MF models assume that users who share similar preferences (i.e. rating positively or negatively the same venues) are likely to prefer similar venues. Such MF models usually treat all users equally, i.e. the predicted rating of a *target* user for a given venue can be influenced by any other user, as long as they share similar preferences, and regardless of their relationship. However, when the users' rating data is sparse in nature, i.e. users/venues have very few ratings, the rating prediction accuracy of traditional MF models can be significantly degraded. To alleviate the sparsity problem, various MF-based approaches have been proposed to incorporate additional information such as the users' friendships (Ma et al., 2011; Guo et al., 2015b; Ma, 2014; Li et al., 2016) and textual content of comments the users leave for venues (Hu et al., 2014; Jin et al., 2016; Chen et al., 2015; Tang et al., 2015). Indeed, those MF-based approaches can be categorised into two categories: namely regularisation techniques and factorisation approaches. In this section, we review both regularisation (Section 3.3.1) and factorisation (Section 3.3.2) approaches that incorporate additional information to improve the accuracy of the rating prediction of traditional MF models.

3.3.1 Social Regularisation Technique

As mentioned in Section 2.1.1.1.2, the traditional regularisation technique based on the Frobenius norm (Equation (2.4)) is commonly used to ensure that the traditional MF models are simple and to avoid over-fitting. Ideally, the traditional regulariser aims to control the magnitudes of the latent factors of users P and venues Q such that P and Q would give a good approximation of user-venue rating matrix R without being too complex. Ideally, the loss of MF in Equation (2.4) will increase if the Frobenius norm of the latent factors is large. Another advantage of the regularisation technique is to alleviate the sparsity problem of MF models. Previous works have shown that when the users' rating data is sparse in nature, i.e. users/venues have very few ratings, the rating prediction accuracy of traditional MF models can be significantly degraded (Ma et al., 2011; Guo et al., 2015b; Hu et al., 2014; Jin et al., 2016). To alleviate the sparsity problem, Ma et al. (2011) proposed a Social Regularisation technique (SoReg) that incorporates social information (e.g. users' friendships) to enhance the prediction accuracy of MF models. They assumed that users are likely to be influenced by their friends who rate similar venues with similar scores. In particular, SoReg aims to minimise the distance between latent factors of target user p_u and his/her friends p_f (i.e.

if the latent factors of two users are close in the latent space, they are likely to get similar recommendations). Ma et al. (2011) introduced their proposed SoReg technique into the traditional MF regularisation (Equation (2.4)) as follow:

$$\underset{SoReg}{L(\Theta)} = \underset{MF}{L(\Theta)} + \frac{\alpha}{2}\sum_{u=1}^{m}\sum_{f\in\mathcal{F}(u)} pcc(u,f)\|p_u - p_f\|_F^2 \tag{3.1}$$

where $\mathcal{F}(.)$ is the set of friends of user u, α is a parameter that controls the influences of the SoReg technique. $pcc()$ estimates the similarity between the ratings of two users using the Pearson Correlation Coefficient (PCC), calculated as follows:

$$pcc(u,f) = \frac{\sum\limits_{i\in V_r(u)\cap V_r(f)} (r_{u,i} - \bar{\mathrm{r}}_u)\cdot(r_{f,i} - \bar{\mathrm{r}}_f)}{\sqrt{\sum\limits_{i\in\mathcal{V}_u^+\cap\mathcal{V}_f^+} (r_{u,i} - \bar{\mathrm{r}}_i)^2}\cdot\sqrt{\sum\limits_{i\in\mathcal{V}_u^+\cap\mathcal{V}_f^+} (r_{f,i} - \bar{\mathrm{r}}_f)^2}} \tag{3.2}$$

where $\mathcal{V}_u^+$ are the venues that user u has rated and $\bar{\mathrm{r}}_u$ is his average rating. This regularisation ensures that friends who have rated venues similarly (e.g. $pcc(u,f) = 0.9$) are predicted to give similar ratings to other venues (i.e. $p_i^T V \approx p_f^T V$). Following Ma et al. (2011), we employ a mapping function $f(x) = \frac{(x+1)}{2}$ to bound the range of the PCC score into [0, 1].

There are three benefits to the SoReg technique. First, similar to the traditional regularisation technique, SoReg can avoid over-fitting and obtaining a factorisation model that is too complex. Second, SoReg ensures the latent factors of similar friends are close to each other in the latent space, whereas latent factors of non-similar friends are allowed to differ. Finally SoReg indirectly models the propagations of tastes (i.e. if user u has a friend f and user f has a friend g, suppose g is not a friend of u, it actually indirectly minimises the distance between the latent factors of p_u and p_g. However, we argue that users are not only influenced by their friends who visit similar venues and provide similar ratings to that venues, but are also influenced by friends who share similar tastes, which can be extracted from the explicit textual comments they have left for each venue. An intuitive scenario for this assumption could be that u, f and g are friends who all enjoy a visit to a restaurant. u and f are impressed by the service and quality of food at the restaurant, but g is only impressed by the setting of the restaurant. Estimating user similarity using Pearson Correlation Coefficient (PCC), as exploited by the SoReg technique (Equation (3.1)), can only capture the similarity of the users based on their ratings (**Limitation M1**). However, in the scenario mentioned above, although u, f and g all enjoy visiting the restaurant and have positive ratings about the restaurant, the taste of g is quite different from the others (i.e. he prefers the setting of the restaurant over the service and food). In particular, the limitation of the SoReg technique can be defined as follows:

Limitation M1: There is a disadvantage of the SoReg technique that still relies on the Pearson Correlation Coefficient to estimate the similarity between users.

Later in Chapter 4, we propose a novel Social and Textual Regularisation technique (STReg) that incorporates both social and textual information to address **Limitation M1**.Moreover, although social information such as the users' friendships can alleviate the sparsity problem and enhance the prediction accuracy of traditional MF models, such social information may not be available due to privacy concerns. In the next section, we describe previous works (e.g. (Hu et al., 2014; Jin et al., 2016)) that improve the quality of the rating prediction of the traditional MF model by incorporating the textual content of comments the users left on the venues they visited.

3.3.2 Textual-based Matrix Factorisation Models

As mentioned in Chapter 1, the explicit textual comments associated with ratings on venues left by users can provide insights about why users rated a given venue positively or negatively, while also reflecting the characteristics of each venue. Indeed, the users' ratings can be influenced by several aspects: namely the characteristics of venues and users' preferences. For example, with respect to the rating scale range from 1 to 5 stars (where 1 star denotes that the user does not like a venue and 5 stars are the most positive ratings), the reason that a user rated a restaurant 4 stars rather than 5 stars could be that he/she was impressed by the food and price but disappointed by the service.

Apart from the social information such as the friendships between users, the explicit textual comments associated with the ratings have been leveraged by previous works to enhance the prediction accuracy of traditional MF models (Hu et al., 2014; Huang et al., 2014; Fu and Li, 2015; Musto et al., 2015; Ozsoy, 2016; Musto et al., 2016; Jin et al., 2016). For instance, Hu et al. (2014) showed that by incorporating the comments left by the users on venues, a more effective rating prediction MF model can be achieved. They assumed that terms that occur in the comments for venue i provide a better description about the venue than the learned latent factors q_i (as obtained in Equation (2.3). To incorporate such the textual contents of comments into the traditional MF model, these terms have to be mapped into the d-dimensional vector space. In particular, inspired by the topic-level decomposition of textual documents proposed by Chen et al. (2012), Hu et al. (2014) decomposed the latent factors of a venue q_i in Equation (2.3) into a combination of latent factors of terms that occurred in the comments of venue i and modified the prediction function of traditional MF model (Equation (2.3)) as follows:

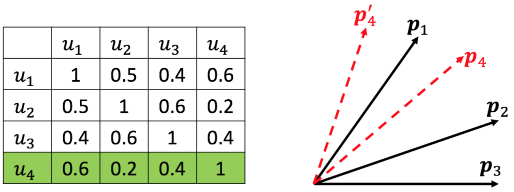
$$\hat{r}_{u,i} = p_u^T \left(\frac{1}{|M_{v_i}|} \sum_{m_{u,i} \in M_{v_i}} \sum_{t \in m_{u,i}} q_{m_t} \right) \tag{3.3}$$

where M_{v_i} is the set of comments the users left for venue i, t is a term that occurs in comment $m_{u,i}$ and $q_{m_t} \in \mathbb{R}^d$ are the latent factors of the comment's term t. Although the Comment-based MF model (CMF) proposed by Hu et al. (2014) can alleviate the sparsity problem for venues that have few ratings by leveraging the textual content of comments, the CMF model lacks flexibility. First, the CMF model requires the latent factors of the comment's terms q_{m_t} to be in the same space as the latent factors of venue q_i (i.e. the dimension d of two latent factors, q_u and q_{m_t}, need to be equal). However, we argue that those two latent factors do not necessarily share the same space due to different nature of venues and comments. Intuitively, similar venues can be recognised by the services provided by those venues, while similar comments can be recognised by terms appearing in the comments and their *semantics*. Therefore, the latent factors of venues and comment's terms should not share the same dimensions, indeed the latent factors of comments should be larger due to the complexity of comments (**Limitation M2**). To alleviate this limitation, Jin et al. (2016) proposed a Joint MF-based approach (JMF) that jointly models the textual content of comment by exploiting the RNN models to encode a sequence of terms in comment $m_{u,i}$ to a d_s-dimensional vector, $s \in \mathbb{R}_s^d$. In particular, given comment $m_{u,i}$, the predicted rating of user u on venue i can be estimated as follows:

$$\hat{r}_{u,i} = p_u^T q_i + p_m^T s_{m_{u,i}} \tag{3.4}$$

where $s_{m_{u,i}}$ is a d_s-dimensional vector that represents comment $m_{u,i}$, which is encoded by the RNN models pre-trained on the large corpus of venues' comments. $p_m \in \mathbb{R}^{d_s}$ is the latent factors of comments that are learnt jointly with the latent factors of user p_u and venue q_i.

The advantage of JMF over CMF is that the latent factors of a venue q_i do not necessarily share the same dimensions as the latent factors of a comment $s_{u,i}$ (i.e. the dimension of latent factors is not equal to the dimension of semantic latent factors, $d \neq d_s$, respectively). Later in Chapter 4, we demonstrate that JMF is more effective than CMF in the rating prediction task. However, we argue that there are two limitations of JMF. First, given rating $r_{u,i}$ and its corresponding comment $m_{u,i}$, JMF aims to capture the preference of user u, the characteristics of venue i and the semantic properties of comment $m_{u,i}$ by jointly learning the latent factors of user p_u, venue q_i and comment p_m. However, instead of leveraging a single comment like the JMF model does, we argue that we can leverage all of the comments that each user left on venues to effectively model his/her preferences. Similarly, we can



	u_1	u_2	u_3	u_4
u_1	1	0.5	0.4	0.6
u_2	0.5	1	0.6	0.2
u_3	0.4	0.6	1	0.4
u_4	0.6	0.2	0.4	1

Figure 3.1: An illustration of the similarity between users and vectors representing the latent factors of each user in user latent space P. Note that these figures are regenerated from He et al. (2017).

also leverage all of the comments for each venue in order to better model its characteristics. Jointly learning a user's preference and the characteristics of a venue from a single comment as the JMF model does may not be effective (**Limitation M3**). In addition, JMF treats the latent factors of user p_u, venue q_i and comment p_m equally, which is counter-intuitive. Indeed, we argue that a more effective model should treat these latent factors independently because they capture different aspects (**Limitation M4**). For example, p_u and q_i capture the user's preferences and the characteristics of venue, while p_m captures the semantic properties of the comments. To address these limitations, in Chapter 4, we propose a novel MF-based approach that leverages all comments to effectively and independently model the users' preferences, the characteristics of venues as well as the semantic properties of comments. In the next section, we describe the state-of-the-art collaborative filtering framework previously proposed in the literature that exploits Deep Neural Networks to effectively model the users' preferences and the characteristics of venues.

3.4 Neural Matrix Factorisation (NeuMF)

In the previous section, we described existing approaches that extend the traditional MF model to incorporate additional information to improve the rating prediction accuracy. Note that these existing MF-based approaches rely on the dot product of the latent factors of user and venue to estimate the rating of user u and venue i, $\hat{r}_{u,i}$. However, He et al. (2017) argued that the dot product of the latent factors may not be sufficient to capture the complex structures of user-venue interactions and can degrade the accuracy of user-venue rating predictions. In the next section, we describe the limitation of the dot product operation used by traditional MF-based approaches and explain what the complex structures of user-venue interactions are.

3.4.1 Limitation of Traditional Matrix Factorisation

Figure 3.1 illustrates a limitation of the dot product operation previously explored by He et al. (2017). The left table in Figure 3.1 shows the cosine similarity between the latent factors of pairs of users. The right vectors in Figure 3.1 represent the latent factors of each user and the geometric relative angles of each user in a 2-dimensional latent factor space. For example, the p_1 and p_2 vectors represent the latent factors of user u_1 and u_2, respectively, and the cosine similarity between these two users is 0.5. Let us first consider users u_1, u_2 and u_3. From the angles of the vectors, we can see that u_1 shares more common preferences with u_2 than u_3 (i.e. indeed the closer the latent factor of two users in the latent factor space, the more similar their preferences are). Next, let us consider u_4, the highlighted row in the table. In fact, u_4 is most similar to u_1, followed by u_3 and u_2. Since u_4 is most similar to u_1, we can place p_4 either on the left or right hand side of p_1 (the two possible geometric positions between p_1 and p_4 in the latent factor space are presented as red-dashed lines). However, either way, this causes p_4 to be closer to p_2 than p_3, which contradicts the fact that u_4 is more similar to u_3 than u_2. This scenario can lead to a large error in Equation (2.4) (i.e. u_4 gets similar venue suggestions to u_2, rather than u_3). Modelling such complex structures of user-venue interactions is a challenging problem.

3.4.2 General framework of NeuMF

He et al. (2017) postulated that the dot product of latent factors used by traditional MF-based approaches may not be sufficient to capture the complex structures of user-venue interactions. To address this challenge, they proposed a Neural Matrix Factorisation Filtering framework (NeuMF) that exploits a Multi-Layer Perceptron (MLP) and non-linear activation functions (described in Section 2.2.1) to model the complex structures of user-venue interactions. In particular, instead of the dot product operation, NeuMF uses the element-wise product and the concatenation of latent factors where the influences of each dimension of the latent factors are captured independently using Deep Neural Networks. Note that, as described in Section 2.1.1.1.2, the dot product operation multiplies and sum the dimensions of the latent factors, which treats each dimension of the latent factors equally, to estimate the rating prediction $\hat{r}_{u,i}$ (see Equation (2.3)).

Figure 3.2 illustrates the multiple layers of the NeuMF framework; the output of one layer serves as the input of the layer above. Starting at the bottom of the figure, the input layer consists of a binary sparse vector with one-hot encoding that represents user u_u and venue v_i, respectively. The sparse vectors of the users and venues are fed into the embedding layer. The outputs of the embedding layer can be seen as the latent factors of user u, $\phi u_u =$

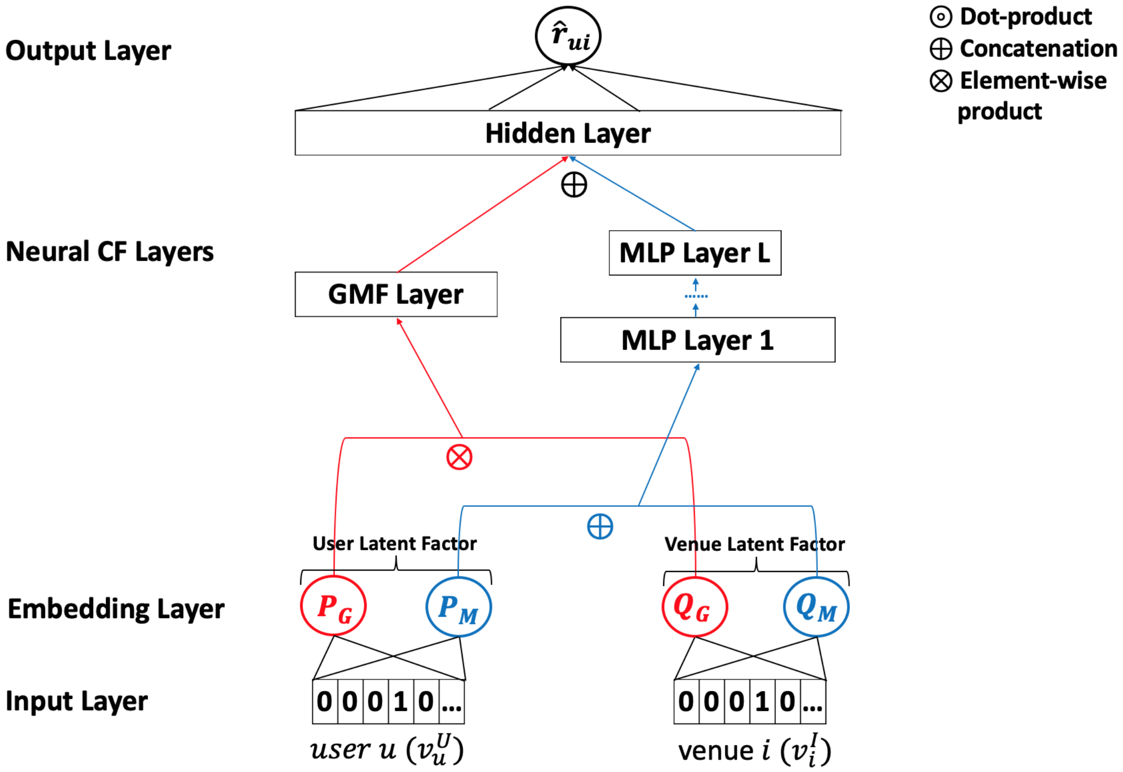


Figure 3.2: The diagram of Neural Matrix Factorisation (NeuMF) framework.

$P^T u_u$, and venue i, $\phi v_i = Q^T v_i$, in the context of factorised model. Next, these latent factors are fed into the Neural Collaborative Filtering layers (i.e. hidden layers) in order to discover latent structures of user-venue interactions. The Neural CF layers consist of two models, namely Generalised Matrix Factorisation (GMF) and Multi-Level Perceptron (MLP), which are further described in Section 3.4.3 and Section 3.4.4, respectively. The connections between layers in the GMF and MLP models are represented using red and blue lines, respectively. The final output layer provides the predicted rating $\hat{r}_{u,i}$, which is defined as follows:

$$\hat{r}_{u,i} = a_{out}(h(\phi^{GMF} \oplus \phi^{MLP})) \tag{3.5}$$

where a_{out} denotes the particular activation function, $\oplus$ denotes the concatenation of the outputs of the GMF and MLP models, ϕ^{GMF} and ϕ^{MLP}, and $h(x) = W^T x + b$ is the hidden layer – W and b are the weight matrix and bias vector, respectively. Overall, $\theta_h = \{W, b\}$ denotes a set of parameters of the hidden layers. $h(x)$ ensures that each dimension of the latent factors from ϕ^{GMF} and ϕ^{MLP} are treated independently. He et al. (2017) proposed to use the sigmoid function $\sigma(x) = 1/(1+e^{-x})$ as the activation function a_{out}. The combination of the GMF and MLP models enables NeuMF to model user-venue interactions as non-linear latent factor models. In particular, the GMF and MLP models capture user-venue interaction using element-wise product and concatenation of latent factors, respectively. Similar to MF's loss function (Equation (2.4)), the NeuMF framework aims to minimise the Root Mean Square Error (RMSE) between the predicted rating $\hat{r}_{u,i}$ and the observed rating $r_{u,i}$.

3.4.3 Generalised Matrix Factorisation model (GMF)

In this section, we describe the first component of the NeuMF framework, a Generalised Matrix Factorisation model (GMF), proposed by He et al. (2017), which aims to mimic the traditional MF model, defined as follows:

$$\phi^{GMF} = \phi u_u^G \otimes \phi v_i^G \tag{3.6}$$

where $\otimes$ denotes the element-wise products of the latent factors of user u and venue i, $\phi u_u^G = P_G^T v_u^U$ and $\phi v_i^G = Q_G^T v_i^I$ are projected from the GMF user and venue embedding layers, $Q_G \in \mathbb{R}^{|\mathcal{V}| \times d}$ and $P_G \in \mathbb{R}^{|\mathcal{U}| \times d}$, respectively [1]. The GMF embedding layers are denoted as the red nodes in the Embedding layer of Figure 3.2. Under the NeuMF framework with GMF model (see Equation (3.5) without ϕ^{MLP}), MF can be easily generalised and extended. Intuitively, if we use an identity function for a_{out} and enforce h to be a uniform

[1] The embedding layer is equivalent to the latent factors of MF-based approaches.

vector of 1 in Equation (3.5), we can exactly recover the traditional MF-based approaches. Moreover, if we allow h to be learned from the element-wise products of two latent factors, ϕu_u^G and ϕv_i^G, without the uniform constraint, it will result in a variant of MF-based approaches that captures the importance of each latent dimensions. Furthermore, by using a non-linear activation function such as the sigmoid function for a_{out} in Equation (3.5), NeuMF can generalise MF-based approaches to a non-linear setting, which may be more expressive than the traditional MF-based approaches that solely rely on the dot product operation. Later in Chapter 5, we extend the GMF model to incorporate the sequence of users' feedback to effectively capture their short-term (*dynamic*) preferences.

3.4.4 Multi-Level Perceptron model (MLP)

In this section, we describe the second component of the NeuMF framework, a Multi-Layer Perceptron model (MLP), which exploits Deep Neural Networks to capture the complex structure of user-venue interactions by using the concatenation of the latent factors. Indeed, the concatenation operation has been widely used in multimodal deep learning work (Srivastava and Salakhutdinov, 2012; Zhang et al., 2014a; Cheng et al., 2016). However, unlike the dot product operation, a vector concatenation of latent factors does not account for any interactions between user and item in a Collaborative Filtering manner. To address this issue, He et al. (2017) proposed to add multiple hidden layers on the concatenated latent factors, using a standard MLP (described in Section 2.2.1) to learn the interaction between the user and venue latent factors, which is defined as follows:

$$\phi^{MLP} = a_L(h_L(...a_1(h_1(\phi u_u^M \oplus \phi v_i^M)))) \tag{3.7}$$

where L is the number of hidden layers, h_x and a_x denote the hidden layer and activation function for the x-th layer's perceptron, respectively. $\phi u_u^M = P_M^T v_u^U$ and $\phi v_i^M = Q_M^T v_i^I$ are the latent factors of user u and venue i that are projected from the MLP user and venue embedding layers, P_M and Q_M, respectively (as illustrated by the blue nodes in the Embedding layer of Figure 3.2). Unlike the GMF model and the traditional MF-based approaches, by concatenating ϕu_u^M and ϕv_i^M, the MLP model is more flexible than GMF and the factorised models since both the dot-product and element-wise product operations require the dimension d of the latent factors to be identical. Regarding the design of hidden layers, a common solution is to follow a pyramid structure, where the hidden layer at the bottom, h_1 is the widest and each successive layer has smaller number of neurons, as illustrated within the MLP layers in Figure 3.2. Indeed, by using a smaller number of hidden units/neurons for the higher layers, they can learn more abstractive features of the latent factors (He et al., 2016a).

Although the activation function a_L can be the sigmoid, hyperbolic tangent (tanh) or Rectified Linear Unit (ReLU), as described in Section 2.2.1, He et al. (2017) argued that ReLU is the most suitable activation function for the hidden layers $h_1...h_L$ for the following reasons. First, the sigmoid function restricts the outputs of each hidden layer between 0 and 1, which may limit the performance of MLP in capturing the complex structure of user-venue interactions. In addition, the sigmoid function usually suffers from the saturation problem, where neurons in the hidden layer stop learning when their output is near either 0 or 1. Although tanh is more preferable than the sigmoid function and widely used in previous literature (Elkahky et al., 2015; Wu et al., 2016), it can be seen as a rescaled version of sigmoid (i.e. $tahn(x/2) = 2\sigma(x) - 1$), which still suffers from the saturation problem. Unlike the sigmoid and $tanh$ functions, Glorot et al. (2011) showed that ReLU can effectively alleviate the saturation problem. Moreover, empirical studies conducted by He et al. (2017) demonstrated that ReLU yields slightly better performance than $tanh$, which in turn is significantly better than sigmoid.

3.4.5 Limitations of NeuMF framework

Experimental results conducted by He et al. (2017) showed that the NeuMF framework is more effective than various traditional MF-based approaches in top-K venue recommendation. However, there are a number of limitations of the NeuMF framework that we need to address in order to generate effective context-aware venue recommendation system. First, as mentioned in Section 3.1, users in LBSNs only focus on the top-K ranked list of venues for obtaining recommendations, and hence we argue that by training of NeuMF while aiming to minimise the Root Mean Square Errors between the predicted ratings and observed ratings may not provide an effective top-K ranked list of venues (**Limitation N1**). Moreover, we argue that both the GMF and MLP models can only capture the users' long-term (*static*) preferences, while previous works (Yu et al., 2016; Zhang et al., 2014b; Tang et al., 2017; Rendle, 2012; Cheng et al., 2013; Koren, 2010a) have shown that users' short-term (*dynamic*) preferences also play an important role in generating effective venue recommendations (**Limitation N2**). Furthermore, although GMF and MLP can capture different structures of user-venue interactions by using both an element-wise product and a concatenation of the latent factors, we argue that the NeuMF framework should not ignore the structure of user-venue interactions that can be captured by the dot-product of latent factors (**Limitation N3**). Finally, as discussed in Section 3.4.2, (He et al., 2017) proposed to apply traditional negative sampling, as defined in BPR (see Algorithm 2.1), to randomly select unvisited venues as negative instances. However, we argue that such a traditional negative sampling approach is not effective and not suitable for CAVR because it does not take the contextual information associated with the observed feedback into account and not all venues are equally likely to be negative instances (**Limitation N4**). We will further discuss this limitation in Section 3.5. Note that **Limitation N1-N3** will be addressed by our proposed Deep Recurrent Collaborative Filtering framework (DRCF), while our proposed *dynamic* geo-based negative sampling approach addresses **Limitation N4**. Our proposed framework and sampling approach are discussed in Chapter 5. Note that these limitations do not only belong to the NeuMF framework and its components but can belong to another approaches. To conclude, in the above analysis, we have identified four limitations of NeuMF and it components as follows:

Limitation N1: There is a disadvantage in the NeuMF framework for identifying the top-ranked venues to present to users as it focuses on rating prediction.

Limitation N2: MF-based approaches for which this limitation applies (GMF, MLP, NeuMF (He et al., 2017)) assume that the users' preferences are *static* and do not account for the sequential properties of observed feedback.

Limitation N3: The MF-based approaches for which this limitation applies (GMF, MLP, NeuMF (He et al., 2017)) ignore the dot product of latent factors that capture user-venue interactions.

Limitation N4: There is a disadvantage in the NeuMF framework that applies the traditional BPR negative sampling approach, in which the contextual information of observed feedback are ignored by the negative sampling approach.

3.5 Negative Sampling Approaches and Ranking Functions for BPR

To generate effective venue recommendations that take the users' context into account, various ranking-based approaches (e.g. BPR proposed by Rendle et al. (2009), discussed in Section 2.1.1.1.2) have been proposed to leverage implicit feedback. As described in Section 2.1.1, Bayesian Personalised Ranking (BPR) is a pairwise ranking-based model that is widely implemented and extended to leverage implicit feedback to generate the top-K venue recommendations (e.g. Yuan et al. (2016); Wang et al. (2016); Zhao et al. (2014); Loni et al. (2016)). Applying the pairwise ranking criterion of the BPR model for venue recommendation assumes that a user prefers visited venues observed in their historical checkins over the non-visited ones. This idea results in a pairwise ranking loss function that tries to discriminate between the set of visited venues and the set of all unvisited venues.

Since users have typically only visited a very small proportion of all venues in LBSNs (Zhang and Chow, 2015; Wang et al., 2016), traditional BPR models typically suffer from the sparsity problem[2] that degrades the quality of the personalised venue suggestions. To enhances the performance of BPR models under such sparsity conditions, various approaches have been proposed in the literature to extend the sampling criterion and pairwise ranking function of BPR to incorporate *additional* sources of information (e.g. social links (Wang et al., 2016; Zhao et al., 2014), geographical information of venues (Yuan et al., 2016) as well as sequential properties of checkins). In the following sections (Section 3.5.1 and Section 3.5.2), we describe the extended BPR models from the literature, which incorporate additional information and identify the limitations of these models. Finally, Section 3.5.3 summarises their elicited limitations. Later, in Chapter 5, we describe our proposed approach that addresses these limitations.

3.5.1 BPR with Geographical Influences

As mentioned in Section 1.1, the geographical information is an important factor that influences the users' decision on visiting novel venues. For example, previous works have shown that users are likely to visit venues nearby their office or home (Cheng et al., 2012; Zhang and Chow, 2015). Yuan et al. (2016) proposed to extend the BPR model to incorporate geographical information (GBPR). In particular, they modified the sampling criterion and pairwise ranking function of BPR based on an assumption that a user is likely to visit a venue g if it is nearby venues that the user has previously visited, $\mathcal{V}_u^+$. Given a user u and a venue they have visited $i \in \mathcal{V}_u^+$, GBPR samples venue g from the *potential feedback* $\mathcal{V}_{u,i}^g$, a set of geographical neighbours of venue i within a μ threshold distance that the user u has not visited before, as a negative example to enhance the effectiveness of the traditional BPR model. In doing so, they proposed a pair ranking function that prefers an unvisited neighbourhood venue $g \in \mathcal{V}_{u,i}^g$ over an unvisited venue $j \in \mathcal{V}_u^-$. The pairwise ranking function of GBPR is defined as follows:

$$\hat{c}_{u,i} \succ \hat{c}_{u,g} \wedge \hat{c}_{u,g} \succ \hat{c}_{u,j}, i \in \mathcal{V}_u^+, g \in \mathcal{V}_{u,i}^g, j \in \mathcal{V}_u^- \tag{3.8}$$

Ideally, the above pairwise ranking function of GBPR aims to rank a visited venue i higher than an unvisited neighbourhood venue g and rank the unvisited neighbourhood venue g higher than an unvisited venue j. Next, for each user $u \in \mathcal{U}$, inspired by the objective function of BPR (Equation (2.7), described in Section 2.1.1.1.3), the objective function of GBPR

[2] A common challenge in recommendation systems when training data is sparse (i.e. users/venues have very few checkins)

can be defined as follows:

$$\mathcal{J}(\Theta) = \underset{\Theta}{argmax} \sum_{u \in \mathcal{U}} \left[\sum_{i \in \mathcal{V}_u^+} \sum_{g \in \mathcal{V}_{u,i}^g} ln(\sigma(\hat{c}_{u,i} - \hat{c}_{u,g})) + \sum_{g \in \mathcal{V}_{u,i}^g} \sum_{j \in \mathcal{V}_u^-} ln(\sigma(\hat{c}_{u,g} - \hat{c}_{u,j})) \right] - \lambda_p \sum_{u \in \mathcal{U}} \|\phi u_u\|_F^2 - \lambda_q \sum_{i \in \mathcal{V}} \|\phi v_i\|_F^2 \tag{3.9}$$

Similar to BPR, GBPR (Yuan et al., 2016) applies Matrix Factorisation to predict $\hat{c}_{u,i}$, the checkin frequency of user u on venue i, by calculating the dot product of the latent factors of user u and venue i, as in Equation (2.3). Finally, Stochastic Gradient Descent (SGD) is applied to find a local maximum of the objective function (Equation (3.9)). Algorithm 3.1 describes an optimisation algorithm of GBPR. In particular, for each iteration (Algorithm 3.1 Lines: 6-8), given a random feedback of user u who has visited venue i, but has not visited neighbourhood venue g and venue j[3], the parameter $\theta \in \Theta$ is updated based on the gradient of its corresponding parameter $\frac{\partial \mathcal{J}}{\partial x}$ while fixing the others, until convergence, as defined in Equation (2.8). Next, the gradients of the latent factor matrices $\phi u_u, \phi v_i, \phi v_g, \phi v_j$ are calculated as follows:

$$\frac{\partial \mathcal{J}}{\partial \phi u_u} = \delta(\hat{c}_{u,i} - \hat{c}_{u,g})(\phi v_i - \phi v_g) + \delta(\hat{c}_{u,g} - \hat{c}_{u,j})(\phi v_g - \phi v_j) - \lambda_p \phi u_u \tag{3.10}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_i} = \delta(\hat{c}_{u,g} - \hat{c}_{u,i})\phi u_u - \lambda_q \phi v_i \tag{3.11}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_g} = \delta(\hat{c}_{u,j} - \hat{c}_{u,g}) - \delta(\hat{c}_{u,g} - \hat{c}_{u,i})\phi u_u - \lambda_q \phi v_g \tag{3.12}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_j} = -\delta(\hat{c}_{u,j} - \hat{c}_{u,g})\phi u_u - \lambda_q \phi v_j \tag{3.13}$$

The advantage of GBPR over the traditional BPR model is that GBPR can leverage the geographical information of venues to effectively sample negative instances based on the pre-defined sampling criterion that users prefer to visit the unvisited neighbourhood venues nearby those venues they previously visited. These are more realistic negative examples than randomly selected distant venues, and hence the effectiveness of the GBPR model is enhanced. Experimental results conducted by (Yuan et al., 2016) showed that GBPR can significantly outperform BPR on several ranking metrics such as Precision and Recall. However, we argue that there are several limitations of GBPR, which are further discussed in Section 3.5.3.

[3] $(u, i, g, j) \in D = \left\{(u, i, g, j) | i \in \mathcal{V}_u^+ \wedge g \in \mathcal{V}_{u,i}^g \wedge j \in \mathcal{V}_u^-\right\}$

Algorithm 3.1: An Optimisation Algorithm of GBPR

```
Input: users $\mathcal{U}$, venues $\mathcal{V}$, visited venues $\mathcal{V}_u^+$, unvisited neighbourhood venue $\mathcal{V}_{u,i}^g$ and
  unvisited venue $\mathcal{V}_u^-$ for each $u \in \mathcal{U}$
Output: $\Theta = \{P \in \mathcal{R}^{m \times d}, Q \in \mathcal{R}^{n \times d}\}$
$P \sim U(0,1), Q \sim U(0,1)$
repeat
    for $u \in \mathcal{U}$ do
        $i \leftarrow$ draw a random visited venue from $\mathcal{V}_u^+$
        $g \leftarrow$ draw random unvisited neighbourhood venue from $\mathcal{V}_{u,i}^g$
        $j \leftarrow$ draw random unvisited venue from $\mathcal{V}_u^-$
        Compute gradients of $\phi u_u$, $\phi v_i$, $\phi v_j$, $\phi v_g$
        Equation (3.10) - (3.13)
        Updated the above parameters
        Equation (2.8)
until convergence;
```

3.5.2 BPR with Social Correlations

Apart from the extended BPR model that incorporates geographical information mentioned in the previous section, there are other works (Wang et al., 2016; Zhao et al., 2014) that have incorporated social information (e.g. social links between users) to sample negative examples based on different criteria. Zhao et al. (2014) proposed a social BPR model (SBPR) that leveraged social links to sample negative examples. Their assumption is that users are likely to visit venues previously visited by their friends. The negative sampling criterion, the ranking function as well as the objective function of SBPR are similar to GBPR's (Equations (3.8) - (3.9)) but substitute $V_{u,i}^g$ with $V_{F_u}^s$ (i.e. a set of venues visited by the user u's friends but which user u has not visited before). In doing so, instead of ranking the unvisited neighbourhood venues higher than the unvisited distant venues, SBPR aims to rank unvisited venues that are previously visited by user's friends higher than venues his/her friends never visited before. Recently, Wang et al. (2016) proposed a finer-grained social BPR that extends SBPR by taking the relationship between friends into account, which is referred to as Strong and Weak-ties (SWBPR): strong-ties are friends who share mutual friends while weak-ties are friends that do not share mutual friends. Their assumption is that venues previously visited by weak-tie friends might be preferred by the user than venues previously visited by strong-tie friends because weak-tie friends are more likely to introduce novel venues. To illustrate their assumption, Wang et al. (2016) assumed that strong-tie friends could be friends from the same high school so they share mutual friends and their preferences are likely to be similar. In contrast, weak-tie friends can introduce new venues that are more interesting. We summarise their proposed ranking criteria as follows:

$$\hat{c}_{u,i} \succ \hat{c}_{u,j}, \text{ if} \begin{cases} i \in \mathcal{V}_u^{+} \wedge j \in \mathcal{V}_u^{joint} & or \\ i \in \mathcal{V}_u^{joint} \wedge j \in \mathcal{V}_u^{weak} & or \\ i \in \mathcal{V}_u^{weak} \wedge j \in \mathcal{V}_u^{strong} & or \\ i \in \mathcal{V}_u^{strong} \wedge j \in \mathcal{V}_u^{none} \end{cases} \tag{3.14}$$

where $\mathcal{V}_u^{joint}$ is the set of venues visited by at least one strong-tie and weak-tie friends of user u, $\mathcal{V}_u^{weak}$ and $\mathcal{V}_u^{strong}$ are the sets of venues visited by at least one of weak-tie friends and strong-tie friends, respectively and $\mathcal{V}_u^{none}$ is the set of venues visited by neither user u nor his/her friends. Note that $\mathcal{V}_u^{joint}$, $\mathcal{V}_u^{weak}$, $\mathcal{V}_u^{strong}$ and $\mathcal{V}_u^{none}$ are all sets of venues that user u has never visited before. Based on their proposed ranking criteria in Equation (3.14), the pairwise ranking function of SWBPR is defined as follows:

$$\begin{gathered} \hat{c}_{u,i} \succ \hat{c}_{u,j} \wedge \hat{c}_{u,j} \succ \hat{c}_{u,w} \wedge \hat{c}_{u,w} \succ \hat{c}_{u,s} \wedge \hat{c}_{u,s} \succ \hat{c}_{u,k}, \\ i \in \mathcal{V}_u^{+}, j \in \mathcal{V}_u^{joint}, w \in \mathcal{V}_u^{weak}, s \in \mathcal{V}_u^{strong}, k \in \mathcal{V}_u^{none} \end{gathered} \tag{3.15}$$

Next, for each user $u \in \mathcal{U}$, the objective function of SWBPR is defined as follows:

$$\begin{aligned} \mathcal{J}(\Theta) = \underset{\Theta}{argmax} \sum_{u \in \mathcal{U}} \Bigg[& \sum_{i \in \mathcal{V}_u^{+}} \sum_{j \in \mathcal{V}_u^{joint}} ln(\sigma(\hat{c}_{u,i} - \hat{c}_{u,j})) \; + \sum_{j \in \mathcal{V}_u^{joint}} \sum_{j \in \mathcal{V}_u^{weak}} ln(\sigma(\hat{c}_{u,j} - \hat{c}_{u,w})) + \\ & \sum_{w \in \mathcal{V}_u^{weak}} \sum_{s \in \mathcal{V}_u^{strong}} ln(\sigma(\hat{c}_{u,w} - \hat{c}_{u,s})) \; + \sum_{s \in \mathcal{V}_u^{strong}} \sum_{k \in \mathcal{V}_u^{none}} ln(\sigma(\hat{c}_{u,s} - \hat{c}_{u,k})) \Bigg] \\ & - \lambda_p \sum_{u \in \mathcal{U}} \|\phi u_u\|_F^2 - \lambda_q \sum_{i \in \mathcal{V}} \|\phi v_i\|_F^2 \end{aligned} \tag{3.16}$$

Similar to BPR and GBPR, SWBPR still relies on the dot product of the latent factors of user u and venue i to predict the checkin of user u on venue i, $\hat{c}_{u,i}$, and Stochastic Gradient Descent is applied to find a local maximum of the objective function (Equation (3.16)). Algorithm 3.2 describes an optimisation algorithm of SWBPR. In particular, for each iteration (Algorithm 3.2, Lines: 6-8), given a random feedback of user u who has visited venue i, but has not visited joint friend venue j, weak-tie friend venue w, strong-tie friend venue s and venue k, the parameter $\theta \in \Theta$ in Equation (3.16) is updated based on the gradient of its corresponding parameter $\frac{\partial \mathcal{J}}{\partial x}$ while fixing the others, until convergence, as defined in Equation (2.8). Next, the gradients of the latent factors $\phi u_u, \phi v_i, \phi v_j, \phi v_w, \phi v_s, \phi v_k$ are calculated as follows:

$$\begin{aligned}\frac{\partial \mathcal{J}}{\partial \phi u_u} =&\delta(\hat{c}_{u,i} - \hat{c}_{u,j})(\phi v_i - \phi v_j) + \delta(\hat{c}_{u,j} - \hat{c}_{u,w})(\phi v_j - \phi v_w) + \\ &\delta(\hat{c}_{u,w} - \hat{c}_{u,s})(\phi v_w - \phi v_s) + \delta(\hat{c}_{u,s} - \hat{c}_{u,k})(\phi v_s - \phi v_k) - \lambda_p \phi u_u\end{aligned} \tag{3.17}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_i} = \delta(\hat{c}_{u,j} - \hat{c}_{u,i})\phi u_u - \lambda_q \phi v_i \tag{3.18}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_j} = \delta(\hat{c}_{u,w} - \hat{c}_{u,j}) - \delta(\hat{c}_{u,j} - \hat{c}_{u,i})\phi u_u - \lambda_q \phi v_j \tag{3.19}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_w} = \delta(\hat{c}_{u,s} - \hat{c}_{u,w}) - \delta(\hat{c}_{u,w} - \hat{c}_{u,j})\phi u_u - \lambda_q \phi v_w \tag{3.20}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_s} = \delta(\hat{c}_{u,k} - \hat{c}_{u,s}) - \delta(\hat{c}_{u,s} - \hat{c}_{u,w})\phi u_u - \lambda_q \phi v_s \tag{3.21}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_k} = -\delta(\hat{c}_{u,k} - \hat{c}_{u,s})\phi u_u - \lambda_q \phi v_k \tag{3.22}$$

Algorithm 3.2: An Optimisation Algorithm of SWBPR

```
Input: users $\mathcal{U}$, venues $\mathcal{V}$, visited venues $\mathcal{V}_u^+$, joint friend venues $\mathcal{V}_u^{joint}$, weak-tie friend
  venues $\mathcal{V}_u^{weak}$, strong-tie friend venues $\mathcal{V}_u^{strong}$ and unvisited venue $\mathcal{V}_u^{none}$ for each
  $u \in \mathcal{U}$
Output: $\Theta = \{P \in \mathcal{R}^{m \times d}, Q \in \mathcal{R}^{n \times d}\}$
$P \sim U(0,1), Q \sim U(0,1)$
repeat
    for $u \in \mathcal{U}$ do
        $i \leftarrow$ draw a random visited venue from $\mathcal{V}_u^+$
        $j \leftarrow$ draw a random joint friend venue from $\mathcal{V}_u^{joint}$
        $w \leftarrow$ draw random weak-tie venue from $\mathcal{V}_u^{weak}$
        $s \leftarrow$ draw random weak-tie venue from $\mathcal{V}_u^{strong}$
        $k \leftarrow$ draw random unvisited venue from $\mathcal{V}_u^{none}$
        Compute gradients of $\phi u_u$, $\phi v_i$, $\phi v_j$, $\phi v_w$, $\phi v_s$, $\phi v_k$
        Equation (3.17 - 3.22)
        Updated the above parameters
        Equation (2.8)
until convergence;
```

The advantage of SBPR and SWBPR over the traditional BPR model is that both SBPR and SWBPR can leverage the users' social information to effectively sample negative instances based on the pre-defined sampling criterion that users prefer to visit the venues previously visited by their friends. Their experiments showed that by incorporating the users' social information during the negative sampling process, more effective ranked list of venue recommendations can be obtained. In the next section, we discuss several limitations of both SBPR and SWBPR models.

3.5.3 Limitations of Existing Negative Sampling Approaches and Ranking Functions for BPR

In this section, we analyse the limitations of the existing negative sampling approaches and the ranking functions for BPR proposed in previous literature (Wang et al., 2016; Zhao et al., 2014; Yuan et al., 2016). Although experimental results obtained by Wang et al. (2016); Zhao et al. (2014); Yuan et al. (2016) all demonstrate that their various proposed negative sampling approaches and ranking functions can significantly improve the performance of BPR in venue recommendation compared to the traditional BPR negative sampling approach proposed by Rendle et al. (2009), there are limitations we need to address to further improve the quality of venue recommendation. First, their proposed negative sampling approaches and ranking function (GBPR, SBPR and SWBPR) are not sufficiently flexible to incorporate multiple types of contextual information (**Limitation S1**). For example, SBPR can only incorporate social information between users but cannot leverage geographical information of venues during the sampling process. In particular, these negative sampling approaches rely on the pre-defined ranking criteria (e.g. SWBPR's ranking criteria in Equation (3.14)), which determines how to compute the gradients of the latent factors. Including additional types of contextual information is relatively difficult because it requires to modify the ranking criteria, the negative sampling approach as well as the gradient computations. Second, the sampling criteria and the ranking function of GBPR, SBPR and SWBPR are based on pre-defined assumptions and not motivated by the characteristics of users' movement and the social interactions in LBSNs that have been observed in previous checkin studies (Cheng et al., 2012; Zhang and Chow, 2015; Zhang et al., 2015a) (**Limitations S2**). We will further discuss this limitation in Chapter 4. Finally, we argue that these negative sampling approaches do not take the sequential order of checkins into account during the sampling approach. However, previous works (Yu et al., 2016; Zhang et al., 2014b; Tang et al., 2017; Zhu et al., 2017; Beutel et al., 2018; Smirnova and Vasile, 2017) have all shown that such the sequential properties of checkins can enhance the quality of venue recommendation (**Limitation S3**). Later in Chapter 4, to address **Limitations S1 & S2**, we describe our proposed Personalised Ranking Framework with Multiple sampling Criteria (PRFMC) that can leverage multiple types of additional information to effectively sample negative examples to enhance the performance of the BPR model. In addition, **Limitation S3** is addressed by our proposed *dynamic* geo-based negative sampling discussed in Chapter 5. Note that these limitations only belong to the negative sampling-based approaches. The limitations of GBPR, SBPR and SWBPR are summarised below:

Limitation S1: The sampling approaches for which this limitation applies (GBPR (Yuan et al., 2016), SBPR (Zhao et al., 2014) and SWBPR (Wang et al., 2016)) are built upon predefined sampling assumptions and are not sufficiently flexible to incorporate different types of additional information.

Limitation S2: The sampling approaches for which this limitation applies (GBPR (Yuan et al., 2016), SBPR (Zhao et al., 2014) and SWBPR (Wang et al., 2016)) is based on predefined assumptions, which are contradicted to the previous studies (Cheng et al., 2012; Zhang and Chow, 2015; Zhang et al., 2015a) that examine users' geographical movements and social influences in LBSNs.

Limitation S3: The sampling approaches for which this limitation applies (GBPR (Yuan et al., 2016), SBPR (Zhao et al., 2014) and SWBPR (Wang et al., 2016)) do not take the sequential order of checkins into account.

Limitations S1 & S2 will be addressed by our proposed Personalised Ranking Framework with Multiple sampling Criteria framework (PRFMC) discussed in Chapter 4. Moreover, **Limitation S3** will be addressed by our proposed sequential-based negative sampling approaches discussed in Chapter 5.

3.6 Recurrent Neural Network Models for Recommendation Systems

Recurrent Neural Networks (RNNs), as described in Section 2.2.3, represent a specific form of DNN models that possess several properties that make them suitable and attractive for sequence modelling (Goodfellow et al., 2016). In particular, RNNs are capable of incorporating input from past interactions, allowing to derive a wide range of sequence-to-sequence mappings. In the past decade, RNNs have been widely applied with considerable success in several domains such as speech recognition, computer vision and natural language processing (e.g. (He et al., 2016a; Zhang et al., 2014a; Kim, 2014)). Recently, various approaches have been proposed to enhance the effectiveness of MF-based approaches for recommendation systems by exploiting RNNs (Yu et al., 2016; Zhang et al., 2014b; Tang et al., 2017; Zhu et al., 2017; Beutel et al., 2018; Smirnova and Vasile, 2017) to leverage the sequential properties of observed implicit feedback. In this section, we describe common techniques that integrate RNN models into MF-based approaches (Section 3.6.1). Then, in Section 3.6.2, we discuss state-of-the-art gating mechanisms for recurrent architectures that take contextual information associated with the sequences of observed checkins (e.g. the time interval and the geographical distance between two successive checkins) into account for CAVR. Finally, Section 3.6.3 summarises the elicited limitations of these RNN-based factorisation approaches and recurrent architectures.

3.6.1 Recurrent-based Factorisation Approaches

Recently, various approaches (e.g. Zhang et al. (2014b); Tang et al. (2017); Yu et al. (2016)) have been proposed to leverage the sequence of users' feedback to capture their *dynamic* preferences by exploiting the traditional RNN models, described in Section 2.2.3.1. Note that these RNN-based approaches have not been originally proposed for venue recommendation but are sufficiently flexible to do so without any disadvantages. For example, Zhang et al. (2014b) proposed a RNN-based factorisation approach (RNN-MF) for click prediction for sponsored search that models users' short-term (*dynamic*) preferences from their sequences of clicks. For simplicity, we explain RNN-MF in the context of venue recommendation as follows:

$$h_\tau = \sigma(X\phi v_i + Rh_{\tau-1}) \tag{3.23}$$

where ϕv_i denotes the latent factor of venue i, the user visited at time step τ and $h_{\tau-1}$ is the *dynamic* preferences of the user at previous time step $\tau - 1$. R is a recurrent connection weight matrix that captures the sequential properties between every two adjacent hidden states $h_{\tau-1}$ and h_τ and X is a transition matrix between the latent factors of venues. Then, similar to the traditional MF-based approaches, they apply the dot product to estimate the probability that user u will checkin at venue i given his recent sequence of checkin $S_{u,\tau}$ (i.e. $\hat{c}_{u,i} = h_\tau \circledcirc \phi v_i$, where $\circledcirc$ denotes the dot product operation). Another example of RNN-based factorisation approach is proposed by Tang et al. (2017). Instead of traditional RNN models, they proposed to exploit a Bidirectional LSTM (BiLSTM) – which is a variation of a LSTM that incorporates the hidden units of LSTM in both forward and backward directions – to model users' short-term preferences from sequences of checkins. Tang et al. (2017) preferred BiLSTM over the traditional RNN models because they assumed that the user's short-term preferences can be effectively captured by considering his/her sequence of checkin in both forward and back directions. Note that both RNN-MF and BiLSTM-MF aim to predict the next checkin and do not take the time of the predicted checkin into account. In particular, their proposed approach (BiLSTM-MF) estimate the probability that user u will checkin at venue i given his recent checkin sequences $S_{u,\tau}$ as follows:

$$\hat{c}_{u,i} = (\phi u_u + h_\tau) \circledcirc \phi v_i \tag{3.24}$$

where h_τ is the output of the BiLSTM and ϕu_u is the latent factors of user u. The advantage of BiLSTM-MF over RNN-MF, apart from the use of more advanced RNN models, is that BiLSTM-MF takes the latent factors of users into account, while RNN-MF does not and solely relies on the user's dynamic preference h_τ captured from the RNN models. The objectives of RNN-MF and BiLSTM-MF are similar to that of MF (Equation (2.4)), in that they aim to minimise the pointwise loss between the predicted checkin $\hat{c}_{u,i}$ and the observed checkin $c_{u,i}$. Moreover, Yu et al. (2016) proposed a Dynamic REcurrent bAsket Model (DREAM), an extension of RNN-MF that incorporates BPR for ranking optimisation. Their experimental results demonstrated that DREAM can significantly outperform various RNN-based approaches including RNN-MF.

We argue that there are three limitations that need to be addressed for RNN-MF, BiLSTM-MF and DREAM. First, RNN-MF does not take the user's long-term (*static*) preferences into account (**Limitation R1**). Although sophisticated RNN-based models such as LSTM and GRU are capable of dealing with long sequences of observed feedback, such models are computationally expensive with respect to the length of sequence. Indeed, a venue that the user has visited a couple of months ago has less impact on the user's short-term (*dynamic*) preferences than a venue recently visited but still has large impact on the user's long-term (*static*) preferences. With long sequence of checkins, LSTM and GRU are not able to capture the large impact of venues the users visited long time ago to model the user's *static* preferences - indeed they likely ignore such historical checkins (**Limitation R1**). Hence, we argue that an accurate model able to capture both the *static* and *dynamic* preferences of users is more likely to generate better venue recommendations. Second, to model the complex structure of user-venue interactions in a Collaborative Filtering manner, RNN-MF and DREAM still rely on a dot product of the latent factors of users and venues $\phi u_u, \phi v_i$ and the user's *dynamic* preferences h_τ. However, as mentioned in Section 3.4, He et al. (2017) showed that the dot product of latent factors may not be sufficient to capture the complex structure of user-venue interactions (**Limitation R2**). However, the operations that combine latent factors $(\phi u_i, \phi v_j)$ and hidden state h_τ in Equation (3.24) need not be limited to the dot product and summation. Finally, we argue that these RNN-based factorisation approaches, RNN-MF, BiLSTM-MF and DREAM, which exploit traditional RNN models to capture the users' dynamic preferences, are not effective, because they only consider the sequence of previously visited venues and ignore the contextual information associated with the checkins (**Limitation R3**). Indeed, the traditional RNN models are not sufficiently flexible to incorporate the contextual information associated with the checkins, which can hinder the performance of modelling the user's *dynamic* preferences. In the next section, we discussed several extensions of the traditional RNN models previously proposed in the literature that aim to address **Limitation R3** by extending the gating mechanism of the RNN architecture to incorporate the contextual information of checkins.

3.6.2 Gating Mechanism of RNN Architectures

As discuss in the previous section, traditional RNN models cannot incorporate the contextual information associated with the sequence of checkins (**Limitation R3**). Consider an illustrative example in Figure 3.3, where user A and B have both visited a museum and a restaurant in the same order. User A visited the museum and restaurant at 9:00 and 18:00, respectively, while User B visited the museum and restaurant at 9:00 and 12:00, respectively. The user's *dynamic* preferences captured by traditional RNN-based models for these two users are similar regardless of the time gaps between successive checkins, hence they are likely to receive similar recommendation. However, we argue that these two users should be presented with different recommendations. Regarding user A, the time interval between checkins at the museum and restaurant is very large, hence the checkin at museum is less likely to influence his/her decision on the next venue to visit. In contrast, the time gap of checkins between the museum and restaurant of user B is shorter, hence the checkin at museum is more likely to influence his/her decision. An example recommendation for these users can be a bar for user A and another museum to visit for venue B. To address **Limitation R3**, various gating mechanisms of RNN architectures have been proposed in previous works (e.g. Zhu et al. (2017); Beutel et al. (2018); Smirnova and Vasile (2017)). In this section, we discuss extensions of the LSTM and GRU architectures [4] proposed by these previous works that aim to incorporate contextual information associated with the sequence of checkins. Although these recurrent architectures were not originally proposed for CAVR, they are sufficiently flexible to be applied to the CAVR task. From now on, we explain their proposed architectures in the context of venue recommendation for reasons of uniformity and use the GRU architecture (see Section 2.2.3.3) to explain their proposed architectures, due to its relative simplicity (i.e. less parameters compared to LSTM).

3.6.2.1 Time-aware GRU architecture (TimeGRU)

As discussed in Section 2.2.3.3, the main advantage of the GRU architecture is that it can alleviate the vanishing gradient problem that is usually suffered by traditional RNN models.

[4] Comprehensive details of LSTM and GRU architectures can be found in Section 2.2.3.

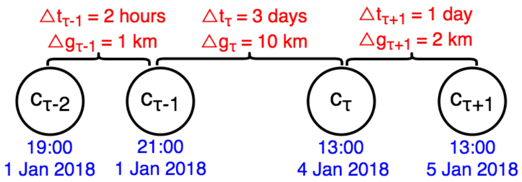

$\Delta t_{\tau-1}$ = 2 hours
$\Delta g_{\tau-1}$ = 1 km
Δt_{τ} = 3 days
Δg_{τ} = 10 km
$\Delta t_{\tau+1}$ = 1 day
$\Delta g_{\tau+1}$ = 2 km
$c_{\tau-2}$
$c_{\tau-1}$
c_{τ}
$c_{\tau+1}$
19:00
1 Jan 2018
21:00
1 Jan 2018
13:00
4 Jan 2018
13:00
5 Jan 2018

Figure 3.3: An illustration of the user's sequence of checkins, where each timestamp of the checkin is highlighted in blue, Δt and Δg are the time interval and the distance between checkins at time step τ, respectively (red text).

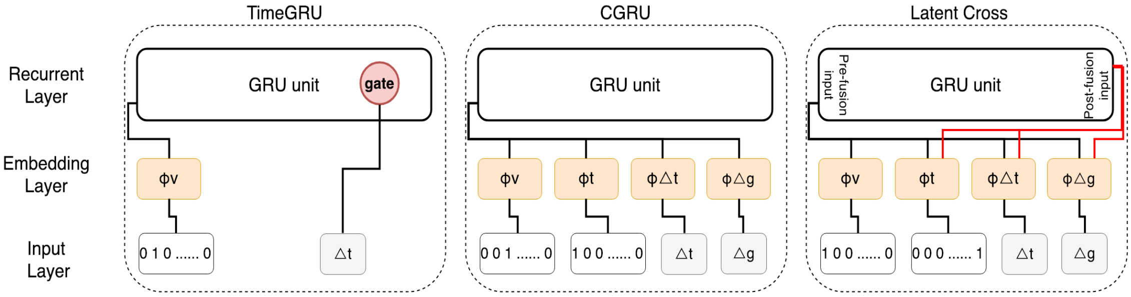


Figure 3.4: Diagrams of existing recurrent architectures for CAVR

However, the GRU architecture is not flexible enough to incorporate the contextual information associated with the sequences of checkins. To address this challenge, Zhu et al. (2017) proposed to extend the GRU units, namely Time-aware GRU architecture (TimeGRU), to incorporate the time interval (i.e. the transition contexts) between successive checkins[5]. The left box of Figure 3.4 illustrates their proposed TimeGRU architecture. In particular, they modified the candidate hidden state $\widetilde{h}_\tau$ of the traditional GRU unit (see Equation (2.31)) with their proposed time gate T_τ, which is defined as:

$$T_\tau = \sigma_t(W\phi v_j^\tau + \sigma(\Delta t_\tau W_t) + b) \tag{3.25}$$

$$\widetilde{h}_\tau = \tanh(W\phi v_j^\tau + R(r_\tau \odot T_\tau \odot h_{\tau-1}) + b) \tag{3.26}$$

where $\Delta t_\tau = t^\tau - t^{\tau-1}$ is the time interval between checkins $s_{i,t}^\tau$ and $s_{i,t}^{\tau-1}$. t^τ captures the correlation between the current venue v_j^τ and the time interval Δt^τ. Intuitively, the time gate T_τ is used to control the influence of previous hidden state $h_{\tau-1}$ in Equation (3.26). In particular, the previous hidden state $h_{\tau-1}$ is not only controlled by the reset gate r_τ but also by their proposed time gate T_τ. Then, to predict the probability that user u will checkin at venue i, given his recent sequence of checkins $S_{u,\tau}$, TimeGRU estimates the predicted checkin $\hat{c}_{u,i}$

[5] Note that although Zhu et al. (2017) used the LSTM architecture to explain their proposed recurrent units, they claimed that their proposed architecture is sufficiently flexible to apply to the GRU architecture.

as follows:

$$\hat{c}_{u,i} = h_\tau \circledcirc \phi v_i \tag{3.27}$$

We argue that there are two limitations of the TimeGRU architecture. First, TimeGRU can only incorporate the time intervals between successive checkins, Δt_τ, (i.e. the *transition* context)) but not the current context of the user, (i.e. the *ordinary* context, such as the time of the day when the user makes a checkin) (**Limitation G1**). Second, their proposed time gate can incorporate only one type of transition context (i.e. the time interval) and is not flexible to incorporate multiple types of transition contexts associated with the checkins, such as adding the geographical distance between two successive checkins (**Limitation G2**).

3.6.2.2 Context-aware GRU architectures

As mentioned in the previous section, one of the limitations of TimeGRU is that it cannot incorporate multiple types of contextual information associated with the sequences of checkins. To address **Limitation G1**, Smirnova and Vasile (2017) proposed a Contextual GRU architecture (CGRU) that can incorporate multiple types of contextual information of served checkins (i.e. both the transition and ordinary context)[6]. Their contributions were two fold: context-dependent venue representations and contextual GRU units. As shown in the second box in Figure 3.4, they proposed a concatenation integration function to model context-dependent venue representations. In particular, at a given time step τ, the input of the GRU unit is the concatenation of the latent factors of the ordinary and transition contexts as well as the latent factors of the venue. Since both the ordinary and transition contexts for the time dimension are continuous values (e.g. the timestamp t^τ, time interval Δt_τ and geographical distance Δg_τ), previous works (Jing and Smola, 2017; Zhao et al., 2016; Smirnova and Vasile, 2017; Beutel et al., 2018) have relied on mapping approaches to represent such context. For example, a source of ordinary context, such as the timestamp t^τ of a venue checkin can be split into discrete features (e.g. month of the year, hour of the day and day of the week). Next, 12, 24 and 7 bits are used to represent the month, hour and day, respectively, and convert to the binary code into a unique decimal digit - a timestamp id. Similarly, the *transition* context (e.g. the time interval Δt^τ) can be quantised into a time interval id using the following function $ind(\Delta t^\tau) = \lceil \frac{\Delta t^\tau}{\delta T} \rceil$, where δT is a 1-hour interval. This technique can be similarly applied to quantise the geographical distance Δg_τ. Then, the timestamp t^τ, the time interval Δt^τ and the geographical distance Δg_τ can be represented as latent factors of time, time interval and distance, $\phi t^\tau, \phi \Delta t_\tau, \phi \Delta g_\tau \in \mathcal{R}^d$, respectively. Next, Smirnova and Vasile (2017) extended the transition matrix W of the GRU unit in Equations (2.29) &

[6] Note that although CGRU was proposed and evaluated in the context of e-commerce item recommendation, it is sufficiently flexible to be applied in the task of CAVR.

(2.31) to be context-dependent, thereby aiming to capture the users' short-term (*dynamic*) contextual preferences. In particular, they introduced the contextual matrix U, to condition the transition matrix W of a GRU unit as follows:

$$\begin{aligned} z_\tau &= \sigma(Wx^\tau \odot U_u xc^\tau) + R_u h_{\tau-1} \\ r_\tau &= \sigma(Wx^\tau \odot U_r xc^\tau) + R_r h_{\tau-1} \\ \widetilde{h}_\tau &= \sigma(Wx^\tau + R_h(r_\tau \odot h_{\tau-1}) \odot U_h xc^\tau) \end{aligned} \tag{3.28}$$

where $x^\tau = [\phi v_i^\tau; \phi t^\tau]$ and $xc^\tau = [\phi t^\tau; \phi \Delta t^\tau; \phi \Delta g_\tau]$ are their proposed context-dependent venue and context representations, respectively.

Recently, building upon the approach of Smirnova and Vasile (2017), Beutel et al. (2018) explored various approaches to effectively incorporate the latent factors of context xc^τ into GRU units. They proposed LatentCross, a technique that incorporates contextual information in GRU, by performing an element-wise product of the latent factors of context xc^τ with the model's hidden states h_τ. The third box in Figure 3.4 illustrates how LatentCross works. The inputs of the GRU unit are the concatenation of all latent factors x^τ (black line) and the concatenation of latent factors of context xc^τ (red line). In particular, they modified Equation (2.32) with the latent factors of context, xc_τ, as follows:

$$h_\tau = (1 + xc^\tau) \odot [(1 - z_\tau)h_{\tau-1} + z_\tau \widetilde{h}_\tau] \tag{3.29}$$

Both CGRU and LatentCross still rely on the dot product between the hidden layer h_τ and the latent factors of users and venues to estimate the predicted checkin $c_{u,i}$, which are similar to RNN-MF and BiLSTM-MF. However, as previously mentioned in Section 3.4.1, He et al. (2017) showed that the dot product operation is not effective to capture the users' preferences and the characteristics of venues from the latent factors of users and venues.

CGRU and LatentCross are the most recent approaches that explore various techniques to incorporate contextual information associated with the sequences of checkins into GRU models. However, we argue that there are two limitations in their proposed GRU architectures. First, both CGRU and LatentCross treat the ordinary and transition context similarly. We argue that different types of context might influence the user's *dynamic* preferences differently (**Limitation G3**). For example, the ordinary context such as time of the day and user's current location should influence the user's contextual preference on the next venue he/she is going to visit, while the transition context such as the time interval between last visited venue and the current time should influence the correlation between the current and previously visited venues. Second, there is a loss of granularity from the quantisation mapping functions used to represent the transition context (**Limitation G4**). For example, instead of representing the timestamp t^{τ} of a venue checkin as a continuous value, the quantisation mapping functions map the timestamp into discrete features (e.g. month of the year, hour of the day and day of the week).

3.6.3 Limitations of RNN models

In this section, we summarise three limitations of RNN-based factorisation approaches (Zhang et al., 2014b; Tang et al., 2017; Yu et al., 2016) and four limitations of the gating mechanism of the GRU architectures (Zhu et al., 2017; Beutel et al., 2018; Smirnova and Vasile, 2017) for CAVR. The limitations of RNN-based approaches are denoted as **Limitation R**, while the limitations of the gating mechanism of the GRU architectures are denoted as **Limitation G**. Note that **Limitation G1-G4** only belong to the gating mechanism of the GRU architectures. The limitations of these approaches and architectures are summarised below:

Limitation R1: The RNN-based factorisation approaches for which this limitation applies (RNN-MF (Zhang et al., 2014b)) do not take the users' long-term (*static*) preferences into account.

Limitation R2: The RNN-based factorisation approaches for which this limitation applies (RNN-MF (Zhang et al., 2014b), BiLSTM-MF (Tang et al., 2017) and DREAM (Yu et al., 2016)) still rely on the dot product operation to combine the latent factors of user ϕu and venues ϕv as well as the hidden unit h_{τ} when predicting a user's checkin.

Limitation R3: There is a disadvantage in the RNN-based factorisation models that model the user's dynamic preferences from sequential order of checkins by leveraging only the sequence of previously visited venues and ignoring the context associated with the checkins.

Limitation G1: The GRU architecture for which this limitation applies (TimeGRU (Zhu et al., 2017)) can only incorporate the transition context (e.g. the time interval between successive checkins) and is not flexible to incorporate the ordinary context (e.g. user's current location).

Limitation G2: The time gating mechanism proposed by Zhu et al. (2017) is not sufficiently flexible to incorporate multiple types of transition contexts associated with the sequence of checkins.

Limitation G3: The GRU architectures for which this limitation applies (CGRU (Smirnova and Vasile, 2017) and LatentCross (Beutel et al., 2018)) treat the ordinary and transition context similarly. As argued in Section 3.6.2.2, these two types of contexts influence the users' preferences differently and should be treated independently.

Table 3.3: Summary of existing works and their limitations

Rating prediction-based approaches					
Model	Additional Info	Context	Sequential	Limitations	Chapter
SoReg	users' social links	×	×	M1	4
CMF	comments	×	×	M2-M4	4
JMF	comments	×	×	M2-M4	4
NeuMF	×	×	×	N1, N4, S3	5
GMF	×	×	×	N2-N3	5
MLP	×	×	×	N2-N3	5
Top-K venue recommendation-based approaches					
Model	Additional info	Context	Sequential	Limitations	Chapter
GBPR	venues' location	×	×	S1-S2	4
SBPR	users' social links	×	×	S1-S2	4
SWBPR	users' social links	×	×	S1-S2	4
RNN-MF	×	×	✓	R1-R3	5
DREAM	×	×	✓	R2-R3	5
Context-aware venue recommendation-based approaches					
Model	Additional info	Context	Sequential	Limitations	Chapter
TimeGRU	×	only time	✓	G1-G4	6
CGRU	×	✓	✓	G3-G4	6
LatentCross	×	✓	✓	G3-G4	6

Limitation G4: There is a disadvantage in the GRU architectures (CGRU (Smirnova and Vasile, 2017) and LatentCross (Beutel et al., 2018)) that rely on the quantised mapping procedures to represent the transition context.

Limitations R1 & R2 will be addressed by our proposed Deep Recurrent Collaborative Filtering framework (DRCF) discussed in Chapter 5. Note that **Limitation R3** has already been addressed by the gating mechanisms of GRU architectures proposed in previous works (e.g. TimeGRU, CGRU and LatentCross). Moreover, **Limitation G1** has already been addressed by the CGRU architecture, while **Limitations G2-G4** will be addressed by our proposed Contextual Attention Recurrent Architecture (CARA) discussed in Chapter 6.

3.7 Roadmap of Addressing Limitations of Previous Works

In this section, we describe a roadmap of this thesis that aims to address the elicited limitations of the previous works. Table 3.3 provides a summary of existing works, their limitations and technical chapters in this thesis that aim to address these limitations. First, regarding **Limitations M1-M4** of the existing rating prediction-based approaches (i.e. SoReg, CMF and JMF), in Chapter 4, we propose several MF-based approaches that aim to address these limitations. In particular, to address **Limitation M1** of the SoReg technique, as mentioned in Section 3.3.1, we propose a novel Social and Textual Regularisation technique (STReg) that can incorporate both the users' social information and the textual content of comments. Unlike the SoReg technique that only relies on the ratings of users' friends to estimate the similarity between the users and their friends, our proposed STReg technique leverages the users' friends' textual information to effectively estimate the similarity between the users and their friends. Moreover, to address **Limitations M2-M4** of the CMF and JMF models, we propose a novel MF-based approach that leverages all comments to effectively and independently model the users' preferences, the characteristics of venues as well as the semantic properties of comments. Furthermore, as discussed in Section 3.5.3, we argue that the existing negative sampling approaches (e.g. GBPR, SBPR and SWBPR) are not sufficiently flexible to incorporate different types of additional information. In particular, to address **Limitations S1-S2** of GBPR, SBPR and SWBPR, in Chapter 4, we propose a novel Personalised Ranking Framework with Multiple sampling Criteria framework (PRFMC) that can incorporate multiple types of additional information to enhance the effectiveness of the traditional BPR model for top-K venue recommendations.

In Section 3.4.5, we have identified various limitations of the existing NeuMF framework. Indeed, we argue that by addressing **Limitations N1-N4** of the NeuMF framework, more effective top-K venue recommendation can be obtained. In Chapter 5, we propose a novel Deep Recurrent Collaborative Filtering framework (DRCF), an extension of the NeuMF framework, which exploits Recurrent Neural Networks to effectively capture the users' *dynamic* preferences from their sequences of checkins. Our proposed DRCF framework consists of three components: namely Generalised Recurrent Matrix Factorisation (GRMF), Multi-Layer Recurrent Perceptron (MLRP) and Recurrent Matrix Factorisation (RMF) models. In addition, within the DRCF framework, we propose novel *dynamic* and *static* geo-based negative sampling approaches that take the sequential properties of checkins and geographical location of venues into account to enhance the effectiveness of the DRCF framework, and to alleviate the cold-start user problem. In particular, the DRCF framework aims to address **Limitations N1-N3**, while our proposed *dynamic* and *static* geo-based negative sampling approaches aim to address **Limitations N4 & S3**. Moreover, within the three components of the DRCF framework, the GRMF and MLRP models aim to address **Limitations R1-R2** of the existing RNN-based factorisation approaches, described in Section 3.6.3.

Next, regarding **Limitations G2-G4** of the existing gating mechanisms of GRU architecture, in Section 3.6.3, we argue that there are different types of contextual information associated with the sequence of checkins (i.e. the *ordinary* and *transition* context). However, these contexts are treated equally by the existing gating mechanisms of GRU architecture. In Chapter 6, we propose a novel Contextual Attention Recurrent Architecture (CARA) for context-aware venue recommendation that effectively incorporates different types of contextual information associated with the users' sequence of checkins. Our proposed CARA architecture consists of two types of gating mechanisms: namely Contextual Attention Gate (CAG) as well as Temporal and Spatial Gates (TSG). The CAG and TSG gates aim to effectively capture the users' contextual preferences from the *ordinary* context associated with the users' checkins and the *transition* context associated with two successive checkins. Finally, in Chapter 7, we combine our proposed DRCF framework with the CARA architecture to effectively generate effective context-aware venue recommendation.

3.8 Conclusions

In this chapter, we first reviewed an extension of the MF-based approach for venue recommendation, NeuMF framework (He et al., 2017), which exploits Deep Neural Networks such as Multi-Layer Perceptron and non-linear activation functions. In particular, in Section 3.4, we discussed the limitations of traditional MF-based approaches that rely on the dot product operation to capture the complex structure of user-venue interactions. Then, we described the NeuMF Framework (Section 3.4.2) as well as its components: namely Generalised Matrix Factorisation (GMF) (Section 3.4.3) and Multi-Layer Perceptron (MLP) (Section 3.4.4) that exploit the element-wise product and concatenation operation, respectively, to capture the complex structure of user-venue interactions. We also identified four limitations of the NeuMF framework (**Limitations N1-N4**) in Section 3.4.5, which we aim to address in Chapter 5. In Section 3.5, we reviewed existing negative sampling approaches and ranking functions for Bayesian Personalised Ranking (BPR) and identified their limitations (**Limitations S1-S3**). Finally, in Section 3.6.1, we reviewed several existing RNN-based factorisation approaches that exploit traditional RNN models capture the users' *dynamic* preferences from the sequential order of checkins. However, such RNN-based factorisation approaches cannot incorporate the contextual information associated with successive checkins due to the limitation of the traditional RNN models. To address this limitation, we described existing gating mechanism of GRU architectures proposed in previous literature (Zhu et al., 2017; Beutel et al., 2018; Smirnova and Vasile, 2017) that can effective incorporate such contextual information. The limitations of RNN-based factorisation and these gating mechanisms of GRU architectures (**Limitations R1-R3** and (**Limitations G1-G4**)) are analysed in Section 3.6.3.

In the next chapter, we describe our proposed STReg technique, a textual-based MF-based approach and Personalised Ranking Framework with Multiple sampling Criteria (PRFMC) that aim to address **Limitations S1 & S2**, **Limitation M1** and **Limitations M2-M4**, respectively.

Chapter 4

Enhancing Collaborative Filtering-based Approaches with Additional Information

4.1 Introduction

In Chapter 2, we provided an overview of basic Collaborative Filtering approaches from the literature such as Matrix Factorisation (MF) (Koren et al., 2009) and Bayesian Personalised Ranking (BPR) (Rendle et al., 2009) that are widely used to leverage the users' observed *implicit* and *explicit* feedback to effectively generate recommendations to users. As we previously discussed in Chapter3 for a LBSN, the users' observed explicit and implicit feedback, such as the ratings and checkins, respectively, are sparse in nature, i.e. users/venues have very few ratings/checkins, which can degrade the accuracy of the rating prediction of the traditional MF models (Ma et al., 2011; Ma, 2014; Jin et al., 2016; Hu et al., 2014) and the quality of the ranked lists of venues made by Bayesian Personalised Ranking (BPR) (Zhang and Chow, 2015; Wang et al., 2016; Yuan et al., 2016; Zhao et al., 2014). To alleviate the sparsity problem and enhance the effectiveness of traditional MF and BPR models, as described in Chapter 3, a common technique is to leverage additional information such as the social information (e.g. users' friendships and the ratings of each user's friends), the geographical information of venues and the textual information (e.g. the textual content of comments associated with the users' rating). To leverage such additional information, various approaches (Ma et al., 2011; Hu et al., 2014; Guo et al., 2015b; Jin et al., 2016) have been proposed to enhance the accuracy of rating prediction of the traditional MF models. Indeed, these approaches can be separated into two categories: namely regularisation techniques and factorisation approaches (see Chapter 3, Sections 3.3.1 & 3.3.2). In addition, to enhance the quality of venue suggestions of BPR, a common approach in previous works (Wang et al., 2016; Zhao et al., 2014; Yuan et al., 2016) is to extend the traditional negative sampling approach and the ranking function of BPR to take those additional information into account (see Chapter 3, Sections 3.5.1 & 3.5.2).

In this chapter, we propose a novel regularisation technique and a textual factorisation approach, which aim to enhance the accuracy of the rating prediction of traditional MF. In particular, our proposed regularisation technique aims to address **Limitation M1** of the Social-based Regularisation (SoReg) technique identified in Section 3.3.1, while our proposed factorisation-based approach aims to address **Limitations M2-M4** of the Comment-based Matrix Factorisation (CMF) and the Joint Matrix Factorisation (JMF) models identified in Section 3.3.2. The description of **Limitations M1-M4** are summarised as follows:

Limitation M1: There is a disadvantage in SoReg in that it still relies on the Pearson Correlation Coefficient to estimate the similarity between users.

Limitation M2: There is a disadvantage in CMF where the dimensions of latent factors of venues and comment's terms are similar.

Limitation M3: There is a disadvantage in JMF for jointly learning a user's preference and the characteristics of a venue from a single comment.

Limitation M4: There is a disadvantage in the CMF and JMF models, which treat different latent factors dependently, although these latent factors capture different aspects.

Moreover, we propose a novel Personalised Ranking Framework with Multiple sampling Criteria framework (PRFMC) that can leverage multiple types of additional information to enhance the effectiveness of the traditional BPR model for top-K venue recommendations. The PRFMC framework aims to address **Limitations S1-S2** of existing extension of the BPR models (i.e. the SBPR[1], SWBPR[2] and GBPR[3] models) identified in Section 3.5.3. The descriptions of **Limitations S1-S2** are summarised as follows:

Limitation S1: The existing BPR models approaches for which this limitation applies (GBPR, SBPR and SWBPR) are built upon pre-defined sampling assumptions and are not sufficiently flexible to incorporate different types of additional information.

Limitation S2: The sampling approaches for which this limitation applies (GBPR,

[1] The Social-based Bayesian Personalised Ranking (SBPR) proposed by Zhao et al. (2014), described in Section 3.5.2

[2] The Strong and Weak social-based Bayesian Personalised Ranking (SWBPR) proposed by Wang et al. (2016), described in Section 3.5.2

[3] The Geo-based Bayesian Personalised Ranking (GBPR) proposed by Yuan et al. (2016), described in Section 3.5.1

SBPR and SWBPR) are based on pre-defined assumptions (e.g. assuming that venues that both the user and his/her friends have never visited can be sampled as negative instances), which are contradicted by the previous studies (Cheng et al., 2012; Zhang and Chow, 2015; Zhang et al., 2015a) that examined the users' geographical movements and the social influences in LBSNs.

The remainder of this chapter is as follows:

- Section 4.2 describes our proposed Social and Textual Regularisation (STReg) technique and our textual factorisation-based approach (MFw2v) that both exploit the word embeddings to capture the semantic properties of comments to improve the prediction accuracy of the traditional MF model. Later, in Section 4.2.4, we evaluate the effectiveness of the STReg technique and that of the MFw2v model for the user-venue rating prediction in comparison with various existing MF-based approaches.
- Section 4.3 describes our proposed Personalised Ranking Framework with Multiple sampling Criteria (PRFMC), an extension of the traditional BPR model that can incorporate multiple sources of additional information to effectively sample negative instances. Section 4.3.6 reports the effectiveness of the PRFMC framework in generating high quality venue recommendations in comparison with state-of-the-art extended BPR models.
- In Section 4.4, we summarise the conclusions of this chapter.

4.2 Enhancing the Rating Prediction of MF with Word Embeddings

As mentioned in Section 3.3, the textual content of comments associated with the users' ratings on venues can provide insights about why the users rated a given venue positively or negatively, while the content of comments also reflects the characteristics of each venue. Hu et al. (2014) and Jin et al. (2016) showed that such comments can be leveraged to enhance the prediction accuracy of the traditional MF model. To incorporate comments into a MF model, typically, the comments are represented using a Bag-of-Words (BoW) approach. However, such a BoW approach may not be effective in capturing the semantic properties of comments, because they ignore the ordering and the semantics of the words. An example of two comments about a venue that are semantically similar are *delicious sushi bar in Illinois* and *best Japanese restaurant in Chicago*. While these two comments have no words in common, their semantic properties are similar. However a similarity measure based upon the BoW representations would fail to capture these properties.

To effectively capture the semantic properties of words, Mikolov et al. (2013a,b) proposed a word embedding technique that represents a term within a multi-dimensional vector space based on the contexts surrounding the term. Word embeddings are being increasingly applied in many applications due to their effectiveness in capturing the semantic properties of textual content, such as text classification (e.g. (Kim, 2014; Yang et al., 2018; McDonald et al., 2017)) as well as recommendation systems (e.g. (Musto et al., 2016; Fu and Li, 2015; Musto et al., 2015)). For example, Musto et al. (2016) applied several word embedding techniques to enhance the effectiveness of content-based collaborative approaches for tasks such as book and movie recommendation. In addition, Fu and Li (2015) showed that using word embeddings to analyse comments not only captures the semantic properties but also the sentiment expressed in the comments.

In this section, we propose a Social and Textual Regularisation (STReg) technique (Section 4.2.1) and a textual-based factorisation approach (MFw2v) (Section 4.2.2) that both exploit word embeddings to capture the semantic properties of the textual content of comments, in order to enhance the accuracy of the rating prediction of a traditional MF model. In particular, our proposed STReg technique aims to address **Limitation M1** of the SoReg technique described in Section 3.3.1, while the MFw2v model aims to address **Limitations M2-M4** of the CMF and JMF models described in Section 3.3.2. Later in Section 4.2.4, we evaluate the effectiveness of our proposed STReg technique and our MFw2v model in enhancing the prediction accuracy of the traditional MF model in comparison with various MF-based baselines.

4.2.1 Social and Textual Regularisation Technique

As mentioned in Section 3.3, although previous works have shown that both social information (Ma et al., 2011) and textual information (Hu et al., 2014; Jin et al., 2016) are important sources of evidence to enhance the accuracy of the rating prediction of the traditional MF model, a model that seamlessly incorporates these two additional sources of information has not been previously proposed. In this section, inspired by Social Regularisation (SoReg) proposed by Ma et al. (2011), which has been described in Section 3.3.1, we propose the novel Social and Textual Regularisation (STReg) technique that extends the traditional MF model by seamlessly incorporating both social information and textual comments by exploiting word embeddings to estimate a semantic similarity of friends based on their explicit textual comments about venues. Similar to the traditional regularisation technique in Equation (2.4), our proposed STReg technique aims to regularise the complexity of the traditional MF model, hence avoiding over-fitting and improving the effectiveness of the MF model. Moreover, as previously discussed in Section 3.3.1, friends who left positive ratings on a venue may have different reasons why they enjoyed visiting the venue (e.g. a user is impressed by the service and quality of food at a restaurant, but his/her friend is only impressed by the setting of the restaurant). Unlike the SoReg technique, our proposed STReg technique aims to address **Limitation M1** by exploiting word embeddings to effectively estimate the semantic similarity of friends based on the textual comments they have left for venues. In particular, following a common technique used by Musto et al. (2015), for a comment $m_{u,i}$ left by user u on venue i, we obtain a word embedding representation of comment $m_{u,i}$ by summing the vectors of the terms that occurred in comment $m_{u,i}$ as follows:

$$w2v(m_{u,i}) = \sum_{t \in m_{u,i}} \vec{\nu}_t \tag{4.1}$$

where t is a term that occurs in the comment $m_{u,i}$ and $\vec{\nu}_t \in \mathbb{R}^k$ is a vector representation of term t obtained from a word embeddings model. k is the number of dimensions in the word embedding space. Thereafter, the similarity between two users based on their comments about venues, which they have both visited, is estimated as:

$$sim_{w2v}(u,f) = \frac{\sum\limits_{i \in \mathcal{V}_m(u) \cap \mathcal{V}_m(f)} sim(w2v(m_{u,i}), w2v(m_{f,i}))}{|\mathcal{V}_m(u) \cap \mathcal{V}_m(f)|} \tag{4.2}$$

where $sim()$ denotes the cosine similarity between two vectors and $\mathcal{V}_m(u)$ is the set of venues for which user u has left a comment. Next, inspired by the SoReg technique proposed by Ma et al. (2011) (Equation (3.1)), we integrate our proposed STReg technique into the traditional MF model as a component of the regularisation (but using comments rather than ratings to identify similar friends):

$$\frac{\alpha}{2} \sum_{u=1}^{m} \sum_{f \in \mathcal{F}(u)} sim_{w2v}(u,f) ||p_u - p_f||_F^2 \tag{4.3}$$

where α is a parameter that controls the influence of the STReg technique. Note that, similar to the traditional regularisation technique in Equation (2.4) and the SoReg technique, our proposed STReg technique can be easily incorporated with various MF-based approaches. For example, we can add our proposed STReg technique (Equation (4.3)) to the loss function of the traditional MF model (Equation (2.4)) and perform Stochastic Gradient Descent (SGD) on latent factors p_u to obtain a local minimum of the loss function while SGD on the latent factor q_i remains the same as in Equation (2.5):

$$\frac{\partial \mathcal{L}_{MF+STReg}}{\partial p_u} = \frac{\partial \mathcal{L}_{MF}}{\partial p_u} + \alpha \sum_{f \in \mathcal{F}(u)} sim_{w2v}(u, f)(p_u - p_f) \tag{4.4}$$

The main advantage of our proposed STReg technique over the SoReg technique for the traditional MF model is that the latent factors of similar friends who have similar tastes, which are extracted from their textual comments instead of the numerical ratings, will be close to each other in the vector space, whereas the latent factors of dissimilar friends will be more different. Later in Section 4.2.4, we empirically evaluate the usefulness of our STReg technique in enhancing the prediction accuracy of the traditional MF model in comparison with the SoReg technique discussed in Section 3.3.1.

As we mentioned earlier in Section 4.1, apart from the regularisation technique, we can also enhance the effectiveness of the traditional MF model for the rating prediction by jointly combining the MF model and word embeddings in a collaborative filtering manner. In the next section, we describe our proposed factorisation-based approach that exploits word embeddings to capture both the semantic preferences of users and the characteristics of venues from textual comments.

4.2.2 Textual-based Matrix Factorisation Approaches

As mentioned in Chapter 3, the users' explicit textual comments can provide insights into why users rate a venue positively or negatively and can also reflect the characteristics of the venues in LBSNs. In Section 3.3.2, we described MF-based approaches previously proposed in the literature, CMF (Hu et al., 2014) and JMF (Jin et al., 2016), which leverage the users' explicit textual comments to model their preferences and to improve the prediction accuracy of the traditional MF model. In this section, we propose a novel MF-based approach (MFw2v) that exploits word embeddings to effectively model the users' preferences and the characteristics of venues from the textual content of comments left by the users in LBSNs. As mentioned in Section 4.1, our proposed MFw2v approach aims to address **Limitations M2-M4** of the Comment-based Matrix Factorisation (CMF) and the Joint Matrix Factorisation (JMF) models, which are identified in Section 3.3.2. We first describe how our proposed MFw2v approach addresses **Limitation M2** of CMF and JMF. In particular, following an approach that is common in the literature (Musto et al., 2016), we exploit the word embeddings to model the user's semantic preferences $s_{u_u} \in \mathbb{R}^k$ and the semantic characteristics of venue $s_{v_i} \in \mathbb{R}^k$ from the comments in a low-dimensional word embedding space, where k is the dimensions of the word embedding vectors, as follows:

$$s_{u_u} = \sum_{m_{u,i} \in M_{u_u}} \sum_{t \in m_{u,i}} w2v(t) \times r_{u,i} \qquad s_{v_i} = \sum_{m_{u,i} \in M_{v_i}} \sum_{t \in m_{u,i}} w2v(t) \times r_{u,i} \tag{4.5}$$

where M_{u_u} and M_{v_i} are the sets of user u's and venue i's comments and $w2v(t) \in \mathbb{R}^k$ is a function that returns a word embedding representation of term t. Note that the $w2v()$ function in Equation (4.5) can be replaced with a word representation generated by more complex Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) (e.g. (Kim, 2014)).

Next, to incorporate the representations of the preferences of each user s_u and the characteristics of venues s_v within the traditional MF model, we introduce the semantic latent factors of users and venues, $P_s \in \mathbb{R}^{m,k}$ and $Q_s \in \mathbb{R}^{n,k}$, respectively. We then modify the prediction function of the MF model in Equation (2.3) as follows:

$$\hat{r}_{u,i} = \alpha p_u^T q_i + (1-\alpha)(p_{s_u}^T s_{u_u} + q_{s_i}^T s_{v_i}) \tag{4.6}$$

where $p_{s_u} \in \mathbb{R}^k$ and $q_{s_i} \in \mathbb{R}^k$ are the semantic latent factors of user u and venue i projected from P_s and Q_s, respectively, and α is a parameter that controls the influence between the latent factors (p_u and q_i) and the semantic latent factors (p_{s_u} and q_{s_i}). Finally, since we update the prediction of the traditional MF model, the local minimum of the loss function of the MF model (Equation (2.4)) based on the Stochastic Gradient Descent (SGD) is re-calculated by optimising each of the latent factors, p_u, q_i, p_{s_u} and q_{s_i}, while fixing the other, until convergence as follows:

$$\frac{\partial \mathcal{L}}{\partial p_u} = \sum_{j=1}^{n} I_{u,i}(r_{u,i} - \hat{r}_{u,i})\alpha q_i + \lambda p_u \qquad \frac{\partial \mathcal{L}}{\partial q_i} = \sum_{i=1}^{m} I_{u,i}(r_{u,i} - \hat{r}_{u,i})\alpha p_u + \lambda q_i \tag{4.7}$$

$$\frac{\partial \mathcal{L}}{\partial p_{s_u}} = \sum_{j=1}^{n} I_{u,i}(r_{u,i} - \hat{r}_{u,i})(1-\alpha) s_{v_i} + \lambda p_{s_u} \qquad \frac{\partial \mathcal{L}}{\partial q_{s_i}} = \sum_{i=1}^{m} I_{i,j}(r_{u,i} - \hat{r}_{u,i})(1-\alpha) s_{u_u} + \lambda q_{s_i} \tag{4.8}$$

$$\frac{\partial \mathcal{L}}{\partial s_{u_u}} = \sum_{j=1}^{n} I_{u,i}(r_{u,i} - \hat{r}_{u,i})(1-\alpha) s_{v_i} + \lambda p_{s_u} \qquad \frac{\partial \mathcal{L}}{\partial s_{v_j}} = \sum_{i=1}^{m} I_{u,i}(r_{u,i} - \hat{r}_{u,i})(1-\alpha) s_{u_u} + \lambda q_{s_i} \tag{4.9}$$

Next, we describe how our proposed MFw2v approach addresses **Limitations M2-M4** of CMF and JMF, which are previously discussed in Section 3.3.2. In particular, there are three main advantages of our MFw2v approach over the CMF and JMF approaches. First, unlike the CMF model (**Limitation M2**), the latent factors of users P and venues Q do not necessarily share the same dimensions as the semantic latent factors of users P_s and venues Q_s (i.e. $d \neq k$). Second, with respect to **Limitation M3**, our proposed MFw2v approach leverages all comments on venues to model the users' semantic preferences and the semantic characteristics of venues, whereas the JMF model only learns to factorise using single comments. Finally, with respect to **Limitation M4**, unlike the JMF model, MFw2v treats all the latent factors P, Q, P_s and Q_s independently, while the influences of these latent factors are controlled by the α parameter.

With respect to our proposed STReg technique, described in Section 4.2.1, MFw2v differs from STReg in several aspects: First, the STReg technique is only beneficial to users who have friends and for LBSNs that record friend relationships. However, the rating prediction of users who do not have any friends will be similar to the rating prediction generated from traditional MF models. In contrast, users who do not have any friends will benefit from the MFw2v approach because (1) users who have similar tastes (i.e. $s_{u_i} \approx s_{u_j}$) are likely to rate venues similarly regardless of their relationship and (2) $q_{s_i}^T s_{v_i}$ ensures that venues that are commented similarly (i.e. $s_{v_i} \approx s_{v_j}$) are likely to share similar semantic characteristics. Second, the MFw2v approach is finer-grained than the STReg technique. In particular, MFw2v learns which dimensions k of the user's semantic preferences s_{u_u} are important via the user's semantic latent factors p_{s_u}. Hence, similar users are likely to have similar semantic latent factors. In contrast, STReg does not learn the semantic latent factors between users but solely relies on the cosine similarity to estimates similarity between two users. Regarding the advantages of the MFw2v approach over the STReg technique discussed above, we expect that MFw2v is more effective than STReg in enhancing the rating prediction of the traditional MF model.

In the next section, we conduct experiments to evaluate the effectiveness of both our proposed STReg technique and our MFw2v approach in enhancing the prediction accuracy of the traditional MF model in comparison with state-or-the-art Social-based Regularisation technique (SoReg) and the existing factorisation-based approaches described in Section 3.3.

4.2.3 Experimental Methodology

In this section, we evaluate the usefulness of our proposed Social and Textual Regularisation technique (STReg) in enhancing the prediction accuracy of the traditional MF model and evaluate the effectiveness of our proposed MFw2v approach in comparison with state-of-the-art MF-based rating prediction approaches. As mentioned in previous section, our proposed STReg aims to address **Limitation M1** of the Social-based Regularisation (SoReg) technique (see Section 3.3.1), while our proposed MFw2v approach aims to address **Limi-**

Table 4.1: Summary of each research question and its corresponding success decision and the limitations of the existing approaches.

Research Question	Limitation	Success Decision
RQ4.1	M1	STReg is more effective than SoReg.
RQ4.2	M2-M4	MFw2v is more effective than CMF and JMF.

tations M2-M4 of the existing textual MF-based approaches: namely the Comment-based Matrix Factorisation (CMF) and the Joint Matrix Factorisation (JMF) (see Section 3.3.2). In particular, we aim to address the following research questions:

- **RQ4.1**: Can we leverage the users' social links and the textual content of comments left by users for venues to enhance the prediction accuracy of the traditional MF model?
- **RQ4.2**: Can we exploit word embeddings to effectively model the user's semantic preferences and the semantic characteristics of venues from the comments to improve the prediction accuracy of the traditional MF model?

Table 4.1 summarises the research questions we aim to address in this section and their corresponding success decision. In particular, to demonstrate that our proposed STReg technique framework can address **Limitation M1** of the SoReg technique, we aim to answer research question RQ4.1 by comparing the performances of STReg and SoReg. Next, by answering research question RQ4.2, we aim to demonstrate the usefulness of word embeddings in modelling the user's semantic preferences and the semantic characteristics of venues from the users' comments.

To answer these research questions, we conduct experiments using the publicly available Yelp dataset[4], which consists of 2,225,214 ratings by 552,339 users for 77,079 venues. It also contains social network information, with $\sim$3.5M friend links. Following common practice (Hu et al., 2014), we remove standard stopwords from comments in the Yelp dataset. We conduct experiments using a 5-fold cross-validation setting, where each fold has 60% training, 20% validation and 20% testing. We implement all experiments using LibRec (Guo et al., 2015a), a Java library for recommendation systems. For each fold, the α parameter in Equation (4.3) and Equation (4.6) is determined using the validation set. Following (Hu et al., 2014; Jin et al., 2016; Ma et al., 2011), we set the dimension of latent factors d to 10 and $\lambda = 0.001$[5]. An experiment conducted by Ma et al. (2011) showed that the value of the

[4] www.yelp.com/dataset_challenge

[5] λ is a parameter that controls the influence of regularisation (see Equation (2.4)).

α parameter[6] has a significant impact on the prediction accuracy of a traditional MF model. Setting α extremely high can degrade the prediction accuracy as users are fully influenced by their friends (i.e. users are predicted to prefer every venue that their friends like). Following Ma et al. (2011)'s approach, we vary $0.000001 \leq \alpha \leq 1$, multiplying α by 10 at each iteration. For each fold, the value of α that minimises Root Mean Square Error (RMSE) on the validation set is used for testing. For word embeddings, we use the Word2Vec tool[7], to train a skip-gram model (Mikolov et al., 2013a) using the default settings (window size 5 and word embedding dimensions $k = 100$) on the Yelp dataset. Previous work by Mikolov et al. (2013b) showed that the skip-gram model performs better than or equally to the Continuous Bag-of-Words model. Finally, we report the user-rating prediction accuracy in term of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) described in Section 2.1.2, which are widely used in the literature (Guo et al., 2015b; Hu et al., 2014; Ma et al., 2011; Jin et al., 2016). Recall that for both MAE and RMSE, lower is better.

Next, we compare our proposed STReg technique and our MFw2v approach with a number of baselines, which can be grouped into 3 categories: traditional Matrix Factorisation model (MF), regularisation techniques and textual MF-based approaches. The baselines are summarised below, while their parameters and sources of evidence are highlighted in Table 4.2.

MF (Koren, 2010b) is the traditional Matrix Factorisation model, which only considers the user-venue matrix to predict the ratings (described in Section 2.1.1.1.2).

SoReg (Ma et al., 2011) is a state-of-the-art social regularisation technique that leverages the users' friendship $\mathcal{F}(.)$ to enhance the prediction accuracy of the traditional MF model (Equation (3.1)), which is described in Section 3.3.1.

BoWReg: Building upon the SoReg technique, BoWReg represents the user's comments using Bag-of-Words (BoW) similarity, and hence considers both the users' friendships and their comments. In particular, this model is similar to our proposed STReg technique but instead the similarity between two users is estimated using an average of the cosine similarity of the user's BoW comment vectors, as follows:

$$sim_{bow}(u,f) = \frac{\sum\limits_{i \in \mathcal{V}_m(u) \cap \mathcal{V}_m(f)} sim(bow(m_{u,i}), bow(m_{f,i}))}{|\mathcal{V}_m(u) \cap \mathcal{V}_m(f)|} \tag{4.10}$$

where $bow(m_{u,i})$ returns a vector that represents the term frequency of terms occurring in comment $m_{u,i}$. Note that the BoW representations are more sparse than the W2V represen-

[6] α is a parameter that controls the influence of the STReg technique (see Equation (4.3))

[7] code.google.com/archive/p/word2vec

Table 4.2: Overview of regularisation techniques and MF-based approaches for the user-venue rating prediction. + indicates an addition to the intuition of the traditional MF model.

Models	Social	Comments	Params	Intuitions
MF	×	×	λ	Users are likely to prefer venues rated that other similar users rate highly.
CMF, JMF	×	✓	λ	+ Users are likely to prefer venues that share similar characteristics (according to textually similar comments).
SoReg	✓	×	λ, α	+ Users are likely to prefer venues that their friends rate highly.
BoWReg, STReg	✓	✓	λ, α	+ Users are likely to prefer venues visited by their friends who have similar tastes.
MFw2v	×	✓	λ, α	+ Users' preferences can be extracted from their comments on venues and users are likely to prefer venues that share similar characteristics.

tation. We use BoWReg as a regularisation baseline that considers both social and textual information without using the word embeddings.

CMF (Hu et al., 2014) is a MF-based approach that leverages the textual content of comments to enhance the effectiveness of the traditional MF model, described in Section 3.3.2. In particular, CMF decomposes the latent factors of a given venue into a combination of latent factors of terms that occurred in the comments of the given venue.

JMF (Jin et al., 2016) is a state-of-the-art rating prediction approach that jointly models comments and user's ratings by exploiting skip-thought vectors (Kiros et al., 2015)[8] to represent the textual content of comments, described in Section 3.3.2. Instead of skip-thought vectors, we re-implement their approach to exploit word embeddings to permit a fair comparison with our proposed MFw2v approach.

[8] a state-of-the-art deep learning approach.

4.2.4 Experimental Results & Discussion

In this section, we report and discuss the experimental results of our both proposed STReg technique and our MFw2v approach in comparison with the baselines mentioned above on the Yelp dataset. In particular, we aim to answer research questions **RQ4.1** and **RQ4.2**. Table 4.3 reports the user-venue rating prediction accuracy, in terms of MAE and RMSE, of our proposed STReg technique and our MFw2v approach as well as other baselines. Firstly, on inspection of our re-implementations of the state-of-the-art regularisation and MF-based baselines in Table 4.3, we note that the relative prediction accuracy of the baselines on the Yelp dataset are consistent with the results reported in (Ma et al., 2011; Jin et al., 2016; Hu et al., 2014), namely that the SoReg technique and the CMF and JMF models outperform the traditional MF model, while JMF outperforms CMF in terms of the MAE and RMSE metrics.

With regard to research question **RQ4.1**, we observe that our proposed STReg technique can enhance the prediction accuracy of the traditional MF model by 7% for MAE ($1.160 \Rightarrow 1.0781$) and 11% for RMSE ($1.5243 \Rightarrow 1.3456$). These results imply that the social information and the textual content of comments are useful in enhancing the accuracy of the rating prediction of the traditional MF model. Moreover, by comparing the STReg and SoReg techniques, the MAE and RMSE results in Table 4.3 show that by addressing the limitation of the SoReg technique (**Limitation M1**, identified in Section 3.3.1, in which SoReg still relies on the Pearson Correlation Coefficient (PCC) to estimate the similarity between users), more accurate rating predictions are obtained. In particular, these results demonstrate that minimising the distance between latent factors of friends based on the semantic similarity of their comments is more effective than estimating user similarity solely based upon similar ratings (i.e. PCC, as per Equation (3.2)). Next, comparing STReg with BoWReg, we find that the prediction accuracy is again enhanced (2% for MAE ($1.1004 \Rightarrow 1.0781$) and 6% for RMSE ($1.4354 \Rightarrow 1.3456$)). These results show that word embeddings are a more effective representation for the textual content of comments than a Bag-of-Words representation. In particular, word embeddings offer a more useful, semantic space for modelling comments by users upon venues that results in a more effective regularisation.

In addressing research question **RQ4.2**, we compare the prediction accuracy of MFw2v in comparison with state-of-the-art textual MF-based baselines. Overall, by addressing the limitations of CMF and JMF approaches (**Limitations M2-M4**)[9], identified in Section 3.3.2, we observe that our proposed MFw2v approach can outperform all textual MF-based baselines, CMF and JMF, in terms of MAE and is as effective as the STReg technique in terms of

[9] The description of **Limitations M2-M4** are summarised in Section 4.1.

Table 4.3: Prediction accuracy in terms of MAE and RMSE of various approaches. Percentage differences of prediction accuracy are calculated with respect to the best performance achieved for that metric, which are highlighted in bold.

Metrics	MF	CMF	JMF	SoReg	BoWReg	STReg	MFw2v
MAE	1.1640	1.2198	1.1795	1.1260	1.1004	1.0781	**1.0188**
Δ	12.47%	16.48%	15.77%	9.52%	7.42%	5.50%	
RMSE	1.5243	1.5006	1.5073	1.3870	1.4354	**1.3456**	1.3458
Δ	11.72%	10.33%	12.02%	2.99%	6.26%		0.01%

RMSE. In particular, comparing with the traditional MF model, the prediction accuracy of MFw2v is $\sim$12% more effective than MF for both MAE and RMSE. In addition, comparing MFw2v with the textual MF-based baselines, CMF and JMF, we find that the prediction accuracy is again improved by approximately 15% and 10% for MAE and RMSE, respectively. These results imply that the users' semantic preferences and the semantic characteristics of venues extracted from the textual content of comments using word embeddings can enhance the rating prediction accuracy of the traditional MF model. Furthermore, note that the textual content of comments of venues are publicly available in LBSNs, while social information such as users' friendships may not be available due to privacy concerns. By comparing our MFw2v approach and STReg technique, the experimental results demonstrate that MFw2v is more effective than STReg in terms of MAE (i.e. 5.5% more accurate than STReg), while it is as effective as STReg in terms of RMSE. Note that the MFw2v approach only takes the users' comments into account, while the STReg technique considers both social information and users' comments. Although the improvements in Table 4.3 are relatively small (e.g. 5.5% improvement in MAE of MFw2v in comparison with STReg), we note that small improvements in MAE and RMSE can lead to marked improvements in the quality of recommendations in practice (Koren, 2010b). Moreover, the observed performances are evaluated over 2.2M ratings. Note that, we expect that the MFw2v approach is more effective than the STReg technique in improving the effectiveness of the traditional MF model due to its advantages over the STReg technique, which are previously discussed in Section 4.2.2.

4.3 Personalised Ranking Framework with Multiple Sampling Criteria

In the previous section, we described our proposed Social and Textual Regularisation (STReg) technique and our textual MF-based approach (MFw2v), which aimed to enhance the effectiveness of the traditional MF model for user-venue rating prediction by exploiting word embeddings to model the textual content of comments. However, as mentioned in Section 3.1, users in LBSNs only focus on the top-K ranked list of venues, while *explicit* rating feedback can be difficult to collect. For these reasons, Bayesian Personalisation Ranking (BPR) (Rendle et al., 2009), which aims to generate accurate ranked lists of venues, leverages *implicit* feedback such as users' checkins, and is considered to be more effective than the rating prediction MF-based models.

A challenge of obtaining implicit feedback from observed checkins by users is that only positive feedback can be observed. Training the MF and BPR models on only positive feedback is not effective because the model is likely to be biased to positive instances. To address this challenge, the negative sampling approach (see Algorithm 2.1) was proposed by Rendle et al. (2009) to uniformly and randomly select venues that the users have not interacted with as negative instances. In Section 3.5, we reviewed several negative sampling approaches from the literature (Wang et al., 2016; Yuan et al., 2016; Zhao et al., 2014), which incorporate additional information such as the users' relationships and geographical information of venues to enhance the effectiveness of the BPR model. However, we argue that these existing negative sampling approaches are not effective due to their limitations (**Limitations S1-S2**), which were identified in Section 3.5.3 and summarised in Section 4.1.

In this section, to address **Limitation S1**, we propose a novel Personalised Ranking Framework with Multiple sampling Criteria (PRFMC) that incorporates multiple types of additional information to further improve the quality of top-K venue recommendation of the BPR model. We first describe the overview of our proposed PRFMC framework for venue recommendation in Section 4.3.1. Then, we describe the two components of the PRFMC framework, which aim to address **Limitation S2** in Sections 4.3.2 & 4.3.3, respectively. Later, in Section 4.3.6, we demonstrate the effectiveness of the PRFMC framework and that of its components in comparison with various state-of-the-art venue recommendation systems.

4.3.1 An Overview of the PRFMC Framework

As mentioned in Section 3.5.3, existing negative sampling approaches (e.g. (Yuan et al., 2016; Wang et al., 2016; Zhao et al., 2014)) are not sufficiently flexible to incorporate different types of additional information (**Limitation S1**). To address this limitation, we propose the novel Personalised Ranking Framework with Multiple sampling Criteria (PRFMC) that can incorporate multiple types of additional information. Inspired by the BPR model, the PRFMC framework consists of a user's preference model, a pairwise ranking function and a negative sampling approach that supports multiple sampling criteria. The overview process of PRFMC framework is described in Algorithm 4.1. Starting with the user's preference modelling, for a given user u and an unvisited venue i, we calculate the user's preference score $p_{u,i}$ based on the product rule as follows:

$$p_{u,i} = \prod_{a \in A} P_a(i|u) \tag{4.11}$$

where A denotes sources of additional information (e.g. the users' friendship and the geographical location of venues) and $P_a(i|u)$ is the estimated probability that user u will visit venue i, which takes a source of additional information a into account. The higher the score, the more likely user u will visit venue i. Note that the product rule has been widely used to fuse different probabilistic models for recommendations in previous works (Zhang and Chow, 2013; Zhang et al., 2015a; Zhang and Chow, 2015; Cheng et al., 2012) and has shown high robustness. Unlike those previous *pointwise* approaches (e.g. (Zhang and Chow, 2015; Zhang et al., 2015a)) that rank venues based on the score $p_{u,i}$ computed in Equation (4.11), we propose to leverage this score to effectively sample negative examples to enhance the effectiveness of the BPR model. The main advantage of the user's preference score $p_{u,i}$ is that it is sufficiently flexible to be extended to incorporate different types of additional information (e.g. social links, geographical information of venues and venue's categories) within A. Indeed, we can include weighting parameters into Equation (4.11) that control the influence of each probabilistic model $P_a(i|u)$. However, following the previous literature (Zhang and Chow, 2013; Zhang et al., 2015a; Zhang and Chow, 2015; Cheng et al., 2012), we use the product operation to combine different probabilistic models. Later in Sections 4.3.2 & 4.3.3, we discuss the probabilistic models that can be combined into Equation (4.11).

Next, for a given checkin $c_{u,i}$, which indicates that the user u visited venue i, we uniformly sample two venues $j, k \in \mathcal{V}_u^-$ that user u has not visited before and then calculate the user's preference scores $p_{u,j}$ and $p_{u,k}$ (see Algorithm 4.1, Lines: 9-11). Then, the pairwise ranking function of the PRFMC framework is defined as follows:

$$\hat{c}_{u,i,j,k}(\Theta) := \begin{cases} \hat{c}_{u,i} \succ \hat{c}_{u,j} \wedge \hat{c}_{u,j} \succ \hat{c}_{u,k}, & \text{if } s_{u,j} > s_{u,k} \\ \hat{c}_{u,i} \succ \hat{c}_{u,k} \wedge \hat{c}_{u,k} \succ \hat{c}_{u,j}, & \text{otherwise} \end{cases} \tag{4.12}$$

Based on the above pairwise ranking function, for each user $u \in \mathcal{U}$, inspired by the objective function of BPR described in Section 2.1.1.1.3, the objective function of PRFMC can be defined as follows:

$$\begin{aligned}\mathcal{J}(\Theta) = \underset{\Theta}{argmax} \sum_{u \in \mathcal{U}} \Bigg[\sum_{i \in \mathcal{V}_u^+} \sum_{j \in \mathcal{V}_u^-} ln(\sigma(\hat{c}_{u,i} - \hat{c}_{u,j})) + \\ \sum_{j \in \mathcal{V}_u^-} \sum_{k \in \mathcal{V}_u^-} ln(\sigma(\hat{c}_{u,j} - \hat{c}_{u,k})) \Bigg] - \\ \lambda_u \sum_{u \in \mathcal{U}} \|\phi u_u\|_F^2 - \lambda_v \sum_{i \in \mathcal{V}} \|\phi v_i\|_F^2\end{aligned} \tag{4.13}$$

where $\sigma(x)$ is the sigmoid function, λ_u, λ_v are regularisation parameters, $\|.\|_F^2$ denotes the Frobenius norm and $\Theta = \{P, Q\}$ denotes all the parameters of the PRFMC framework to be learnt. Similar to the BPR, SBPR, SWBPR and GBPR models, we apply Matrix Factorisation to predict the checkin frequency of user u on venue i by calculating the dot product of the latent factors of user u and venue i (i.e. $\hat{c}_{u,i} = \phi u_u^T \phi v_i$). Note that our proposed PRFMC framework is sufficiently flexible to use more-sophisticated MF-based checkin prediction approaches for estimating $\hat{c}_{u,i}$ in Equation (4.13). For example, instead of the dot product operation, we can apply the prediction function of the NeuMF framework that relies on the concatenation and element-wise product of latent factors of users and venues (see Equation (3.5), page 46). However, for simplicity, we apply the traditional MF model to predict the users' checkins. Finally, we use Stochastic Gradient Descent (SGD) to find a local maximum of the objective function (Equation (4.13)). In particular, for each iteration (Algorithm 4.1, Lines 13-15), given a random feedback tuple of user u who has visited venue i, but not visited venues j and k, $(u, i, j, k) \in D = \{(u, i, j, k) | i \in \mathcal{V}_u^+ \wedge j, k \in \mathcal{V}_u^-\}$, we update the model parameter $\theta \in \Theta$ based on the gradient of its corresponding parameter $\frac{\partial \mathcal{J}}{\partial x}$ while fixing the others, $\Theta - \theta$, until convergence, as follows:

$$\theta^{(\mathcal{T}+1)} = \theta^{(\mathcal{T})} + \eta^{(\mathcal{T})} \cdot \frac{\partial \mathcal{J}}{\partial \theta}(\theta^{(\mathcal{T})}) \tag{4.14}$$

For the purpose of SGD, the gradients of the latent factors ϕu_u, ϕv_i, , ϕv_j, ϕv_k are calculated as follows:

$$\frac{\partial \mathcal{J}}{\partial \phi u_u} = \delta(\hat{c}_{u,j} - \hat{c}_{u,i})(\phi v_i - \phi v_j) + \delta(\hat{c}_{u,k} - \hat{c}_{u,j})(\phi v_j - \phi v_k) - \lambda_u \phi u_u \tag{4.15}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_i} = \delta(\hat{c}_{u,j} - \hat{c}_{u,i}) \phi u_u - \lambda_v \phi v_i \tag{4.16}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_j} = (\delta(\hat{c}_{u,k} - \hat{c}_{u,j}) - \delta(\hat{c}_{u,j} - \hat{c}_{u,i})) \phi u_u - \lambda_v \phi v_j \tag{4.17}$$

$$\frac{\partial \mathcal{J}}{\partial \phi v_k} = -\delta(\hat{c}_{u,k} - \hat{c}_{u,j})\phi u_u - \lambda_v \phi v_k \tag{4.18}$$

Algorithm 4.1: An Optimisation Algorithm for PRFMC

```
Input: users $\mathcal{U}$, venues $\mathcal{V}$, visited venues $\mathcal{V}_u^+$ and social links $F_u$ for each $u \in \mathcal{U}$
Output: $\Theta = \{P \in \mathcal{R}^{m \times d}, Q \in \mathcal{R}^{n \times d}\}$
$P \sim U(0,1), Q \sim U(0,1)$ $\mathcal{T} \leftarrow 0$ // iteration number
repeat
    for $\mathcal{T} \leftarrow 1$ to $|\mathcal{U}|$ do
        $u \leftarrow$ draw a random user from $\mathcal{U}$
        $i \leftarrow$ draw a random visited venue from $\mathcal{V}_u^+$
        $j, k \leftarrow$ draw random unvisited venues from $\mathcal{V}_u^-$
        if $p_{u,k} > p_{u,j}$ then
            swap $j$ and $k$
        Compute gradients of $\phi u_u$, $\phi v_i$, $\phi v_j$, $\phi v_k$
        Equations (4.15 - 4.18)
        Update the parameters ($\phi u_u$, $\phi v_i$, $\phi v_j$, $\phi v_k$)
        Equation (4.14)
until convergence;
```

There are two main advantages of our proposed PRFMC framework over the SBPR, SWBPR and GBPR models. First, the PRFMC framework is more flexible than those existing BPR-based models in incorporating additional sources of information. Note that, to incorporate additional information within the SBPR, SWBPR and GBPR models, we need to (1) adjust their sampling criterion, then (2) adjust their pairwise ranking function and (3) recalculate the gradients of the latent factors of users and venues. However, with PRFMC, we only need to extend the user's preference score function $s_{u,i}$ in Equation (4.11) to incorporate additional probabilistic models. In particular, we can simply add a new probability model $P_a(i|u)$ that takes the new additional source of information a into account in Equation (4.11). In addition, unlike SBPR, SWBPR and GBPR, our proposed PRFMC framework does not need to revise the gradient calculations when the new additional sources of information are introduced to the framework (i.e. Equations (4.15) - (4.18) remain the same). However, note that, this advantage of PRFMC no longer applies if more advanced DNN libraries such as Tensorflow[10] and Keras[11] are deployed because these libraries support automatic gradient calculations.

Another advantage of the PRFMC framework is its efficiency. The computational complexity of our proposed PRFMC framework consists of the calculation of the rating prediction from the traditional MF model, the pairwise learning algorithm as well as the preference

[10] https://www.tensorflow.org/

[11] https://keras.io/

score function (Equation (4.11)). In particular, the training time of the traditional MF model scales linearly with the number of checkins in C (Koren et al., 2009). Regarding the complexity of our proposed pairwise learning algorithm, the computation of each gradient is $O(d)$ (Equations (4.15)-(4.18)), where d is the dimensions of latent factors. Note that the probabilistic models, which are described in Sections 4.3.2 & 4.3.3, can be pre-computed, hence the complexity of the scoring function at training and testing time is $O(1)$. Overall, the total complexity of PRFMC is $O(\mathcal{T} \cdot |\mathcal{U}| \cdot d)$, where $\mathcal{T}$ is the number of iterations and $|\mathcal{U}|$ is the number of users. Indeed, the computational complexity of our proposed PRFMC framework is equivalent to the BPR, GBPR, SBPR and SWBPR models. Therefore, the efficiency and scalability to large datasets of the PRFMC framework is similar to those of the BPR-based models. In the next section, we describe how to integrate state-of-the-art probabilistic models into the PRFMC framework to effectively sample negative instances.

4.3.2 A Negative Sampling Criterion with Geographical Influence

As discussed in Section 3.5.1, Yuan et al. (2016) proposed an extension of the BPR model (GBPR) that leverages the geographical information of venues to sample negative examples from unvisited venues nearby a previously visited venue i, $\mathcal{V}^{g}_{u,i}$. As discussed in Section 3.5.3, we argue that the negative sampling approach of GBPR is not effective. In particular, GBPR's sampling criterion is contradictory to previous studies (Cheng et al., 2012; Ye et al., 2011; Yuan et al., 2013; Zhang and Chow, 2013; Zhang et al., 2015a; Zhang and Chow, 2015) that examined users' geographical movements on LBSNs. In particular, these previous works have shown that users in LBSNs are likely to visit new venues nearby to venues that they often visit (e.g. their home, office and travel places). However, GBPR's sampling criterion uniformly samples negative venues nearby to any previously visited venues, regardless of how often they have visited other venues in the same area. To illustrate the users' geographical movements in LBSNs observed in previous studies, Figure 4.1 shows the checkin characteristics of a particular user in different cities (centres) of the USA. In centre 1 (recently visited area), the user has only visited one venue, while he/she has visited various venues in centre 2 (his/her home area). Hence, the user is more likely to visit venues nearby to venues in his/her home area (centre 2) rather than the recently visited area (centre 1). However, we argue that the negative sampling criterion of GBPR, which ignores the users' geographical movements – as widely explored in previous literature – can lead to a non-optimal sampling approach (**Limitation S2**).

To address **Limitation S2** of the GBPR model, we propose a novel sampling criterion that takes the users' geographical movements into account, which are captured by a geo-

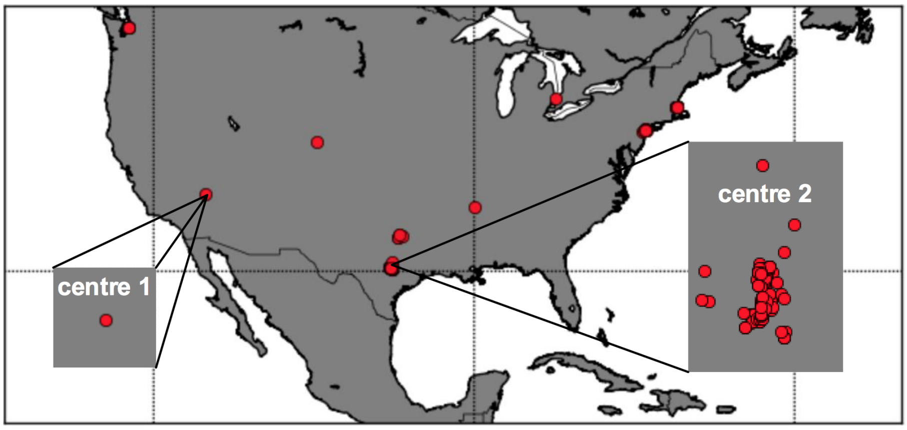


Figure 4.1: A typical user's multi-centres checkin behaviour sampled from the Brightkite dataset. This figure is obtained from Manotumruksa et al. (2017b).

graphical probabilistic model. Note that, previous works (e.g. (Cheng et al., 2012; Zhang and Chow, 2015; Zhang et al., 2015a)) ranked venues based on the probabilistic scores generated by the geographical probabilistic model. In contrast, we use the geographical probabilistic model to estimate the preference score $p_{u,i}$ in Equation (4.11) to effectively sample negative instances, instead of ranking venues. Later in Section 4.2.4, we demonstrate that exploiting the probabilistic scores to sample negative instances is more effective than using the scores to rank unvisited venues in enhancing the quality of top-K venue recommendation. In addition, we demonstrate that the PRFMC framework with the geographical probabilistic model significantly outperforms GBPR.

To effectively capture the users' geographical movements on LBSNs, we apply the Multi-centre Gaussian model (MGM) proposed by Cheng et al. (2012) to calculate the probability of a user u, visiting venue i, given a pre multi-centre of the user $\mathbb{C}_u$, $P_m(i|\mathbb{C}_u)$, (i.e. a pre-calculated list of frequently visited areas/centres) as follows:

$$P_m(i|\mathbb{C}_u) = \sum_{c \in \mathbb{C}_u} P(i \in c) \frac{f_c^\alpha}{\sum_{j \in \mathbb{C}_u} f_j^\alpha} \frac{\mathcal{N}(i|\mu_c, \sigma_c)}{\sum_{j \in \mathbb{C}_u} \mathcal{N}(i|\mu_j, \sigma_j)} \tag{4.19}$$

Equation (4.19) consists of a marginalisation of the product of three terms, namely:

- $P(i \in c)$, $\propto 1/dist(i, c)$, is inversely proportional to the distance between venue i and the centre c, $dist(i, c)$.
- $\frac{f_c^\alpha}{\sum_{j \in \mathbb{C}_u} f_j^\alpha}$ denotes the normalised effect of checkin frequency f_c in the centre c, where $\alpha \in (0, 1]$ controls the checkin frequency property (i.e. the smaller α is the less significant effect on the checkin frequency).

- $\frac{\mathcal{N}(i|\mu_c,\sigma_c)}{\sum_{j\in\mathbb{C}_u}\mathcal{N}(i|\mu_j,\sigma_j)}$ denotes the probability of a venue belonging to the centre c, where $\mathcal{N}(i|\mu_c,\sigma_c)$ is the probability density function of a Gaussian distribution, while μ_c and σ_c correspond to the mean and covariance distances of venues located around the centre c.

Next, we use a greedy clustering algorithm proposed by Cheng et al. (2012) to find the multi-centres of a user u, $\mathbb{C}_u$. For each user u, we start from the most visited venue of the user in $\mathcal{V}_u^+$, and combine all other visited venues from $\mathcal{V}_u^+$ whose distance is less than κ kilometres from the selected venue, into a given region. If the ratio of the total checkin number of venues in this region to the user's total checkin number is greater than a threshold ϕ, we set these checkin venues as a region and determine the most visited checkin venue as the centre of the region. Algorithm 4.2 shows the procedure for discovering the multiple centres of all users. In this section, we have described how to integrate the geographical probabilistic model into the PRFMC framework to effectively sample negative instances. In particular, by applying the Multi-centre Gaussian model (MGM), the PRFMC framework can take the users' geographical movements (e.g. users are likely to visit venues nearby their centres) into account during the sampling process, which is more effective than the negative sampling approach of the GBPR model. As mentioned in Section 4.1, the existing extension of the BPR models (e.g. the SBPR, SWBPR and GBPR models) can only leverage one type of additional information. In the next section, we describe how to integrate a state-of-the-art social probabilistic model to the PRFMC framework to leverage different types of additional information in order to effectively sample negative instances.

Algorithm 4.2: Multi-centre Discovering Algorithm (Cheng et al., 2012)

```
for $u \in \mathcal{U}$ do
    Sort all venues in $\mathcal{V}_u^+$ according to visiting frequency
    $\forall i \in \mathcal{V}_u^+,\ v_i.centre = -1$
    $centre\_list = \emptyset, centre\_no = 0$
    for $i = 1 \rightarrow |\mathcal{V}_u^+|$ do
        $centre\_no{+}{+},\ centre = \emptyset$
        $centre.total\_freq = 0$
        $centre.add(v_i), centre.total\_freq \mathrel{+}= v_i.freq$
        for $j = i + 1 \rightarrow |V_u^+|$ do
            if $v_i.centre == -1 \wedge dist(v_i, v_j) \leq \kappa$ then
                $v_j.centre = centre\_no, centre.add(v_j)$
                $centre.total\_freq \mathrel{+}= v_j.freq$
        if $centre.total\_freq \geq u.total\_freq \times \phi$ then
            $centre\_list.add(centre)$
    return centre_list for u
```

4.3.3 A Negative Sampling Criterion with Social Correlation

In Section 4.2.4, our experimental results showed that our proposed Social and Textual Regularisation technique (STReg) that leverages the users' social information can enhance the effectiveness of the traditional MF model. Similarly, previous literature (Wang et al., 2016; Zhang and Chow, 2015; Zhao et al., 2014; Zhang et al., 2015a; Cheng et al., 2012) have shown that social influences play an important role in venue recommendation systems because users are likely to visit similar venues their friends visited. For instance, Zhang and Chow (2015) shown that users are more likely to visit venues that their friends often visited and similarly friends are also likely to visit similar venues and such social influences follow the power-law distribution.

However, as mentioned in Section 3.5.3, we argue that the negative sampling approach of SBPR and SWBPR models do not take the social influences previously observed in previous studies into account, which may lead to a non-optimal negative sampling approach (**Limitation S2**). Indeed, the sampling criterion of SBPR and SWBPR are based on their proposed pre-defined sampling assumptions, which are contradictory to previous studies (Zhang and Chow, 2015; Zhang et al., 2015a) about social influences on LBSNs. For instance, the negative sampling approach of SBPR uniformly sample venues a given user has not visited before but have been visited by the user's friends and negative instances, regardless of the degree of their friendship (e.g. how many similar venues friends have visited and how often they have visited those venues).

To address **Limitation S2** of SBPR and SWBPR models, we propose to integrate the social relevance model based on the power-law distribution proposed by Zhang and Chow (2015) into the PRFMC framework. By doing so, our proposed PRFMC framework will take the users' social information into account during the negative sampling process. Again, we note that our contribution in this section differs from that of Zhang and Chow (2015). Indeed, we apply the social relevance model to effectively sample negative examples, while they used the probabilistic scores generated by the social relevance model to rank unvisited venues. Later in Section 4.2.4, we demonstrate that the social relevance scores to sample negative instances are more effective than using the social relevance scores to rank unvisited venue in enhancing the quality of top-K venue recommendation. In addition, we demonstrate that the PRFMC framework with the social relevance model significantly outperforms several social-based BPR approaches (SBPR and SWBPR).

The social relevance model proposed by Zhang and Chow (2015) consists of three steps: social aggregation, distribution estimation of social checkin frequency and social relevance score computation.

Step 1: Social aggregation. Given a user u and an unvisited venue i, we aggregate the checkin frequency of the user's friends F_u on venue i, as follows:

$$x_{u,i} = \sum_{f \in F_u} c_{f,i} \tag{4.20}$$

Then we transform the social checkin frequency into normalised relevance based on the social checkin frequency distribution, which is learned from the historical checkin of all users.

Step 2: Distribution estimation of social frequency. In real-world datasets, the social checkin frequency random variable x follows a power-law distribution (Zhang and Chow, 2015), the probability density function of which is defined by:

$$f_{So}(x) = (\beta - 1)(1 + x)^{-\beta},\ x \geq 0,\ \beta > 1 \tag{4.21}$$

where β is estimated by the checkin matrix R and the social links matrix F, as follows:

$$\beta = 1 + |\mathcal{U}||\mathcal{V}| \left[\sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{V}} ln(1 + \sum_{f \in F_u} c_{f,i}) \right]^{-1} \tag{4.22}$$

Step 3: Social relevance score computation. The estimated probability density function f_{So} in Equation (4.21) is monotonically decreasing with respect to the social checkin frequency x, but the social relevance score should be monotonically increasing with regard to the social checkin frequency, because users who have friends with whom they share common visited venues should have high social relevance scores. Thus, we define the social relevance score of $x_{u,l}$ in Equation (4.20) based on the cumulative distribution function of f_{So}, given by:

$$P_s(i|u) = \int_0^{x_{u,i}} f_{So}(z)dz = 1 - (1 + x_{u,i})^{1-\beta} \tag{4.23}$$

such that $P(i|u)$ is monotonically increasing with respect to the social checkin frequency $x_{u,i}$. Moreover, based on the cumulative distribution probability $P(i|u)$ in Equation (4.23), the social checkin frequency $x_{u,i}$ is transformed into a social relevance score that reflects the relative position of $x_{u,i}$ in all the social checkin frequencies of users on venues. Finally, by adding the social relevance model ($P_s(i|u)$ in Equation (4.21)) and the geographical probabilistic model ($P_m(i|\mathbb{C}_u)$ in Equation (4.19) into the user's preference model in Equation (4.11), our proposed PRFMC framework can leverage both users' geographical movements and social influences to effectively sample negative instances. We therefore argue that the PRFMC framework is more flexible than the SBPR, SWBPR and GBPR models in incorporating multiple sources of additional information. In the next section, we evaluate the effectiveness of the PRFMC framework in comparison with various state-of-the-art BPR-based approaches.

4.3.4 Experimental Methodology

In the previous section, we described our proposed Personalised Ranking Framework with Multiple sampling Criteria (PRFMC), which aims to address the limitations of existing BPR-based approaches (**Limitations S1 & S2**). In this section, we evaluate the effectiveness of our proposed PRFMC framework for top-K venue recommendations in comparison with state-of-the-art BPR-based baselines, namely the SBPR, SWBPR and GBPR models. Note that **Limitation S1** of the baselines (i.e. they are not sufficiently flexible to incorporate different types of additional information) has been addressed in the PRFMC framework and does not require experimental verification because PRFMC is sufficiently flexible to incorporate different types of additional information (see Section 4.3.1). In particular, we aim to answer the following three research questions:

RQ4.3 *Can we effectively sample negative venues by leveraging the users' geographical movements previously observed in the literature?*

RQ4.4 *Can we effectively sample negative venues by leveraging the social influences previously observed in the literature?*

RQ4.5 *Is a negative sampling approach based on multiple criteria more effective than a sampling approach with a single criterion in improving the quality of top-K venue recommendations?*

Table 4.4 summarises the research questions we aim to address in this section and their corresponding success decision. In particular, to demonstrate that our proposed PRFMC framework can address **Limitation S2** of the existing social-based negative sampling approaches (i.e. SBPR and SWBPR), we aim to answer research question RQ4.3 by comparing the performances of the SBPR and SWBPR models with our proposed PRFMC framework that incorporates the SPLD model during the sampling process. Next, by answering research question RQ4.4, we aim to demonstrate whether the PRFMC framework that incorporates the MGM model during the sampling process can address **Limitation S2** of the geo-based negative sampling approach (i.e. GBPR). Finally, by answering research question RQ4.5, we can demonstrate the usefulness of the multiple sampling criteria in enhancing the quality of top-K venue recommendations.

Table 4.4: Summary of each research question and its corresponding success decision and the limitations of the existing approaches.

Research Question	Limitation	Success Decision
RQ4.3	S2	PRFMC with the Social Power-Law Distribution model (SPLD) is more effective than the existing social-based negative sampling approaches.
RQ4.4	S2	PRFMC with the Multi-centre Gaussian Model (MGM) is more effective than the existing geo-based negative sampling approaches.
RQ4.5	S2	PRFMC with both SPLD and MGM is more effective than social- and geo-based negative sampling approaches.

To answer these research questions, we conduct several experiments using publicly available large-scale LBSN datasets. In particular, to show the generalisation of our proposed PRFMC framework across multiple LBSN platforms and sources of feedback evidence, we use two checkins datasets from Gowalla and Brightkite[12], and a rating dataset from Yelp[13]. For each dataset, we conduct experiments using a 5-fold cross-validation, where each fold has 60% training, 20% validation and 20% test instances (checkins/ratings). Due to the high sparsity of the datasets, we follow the common practice from previous works (e.g. (Rendle et al., 2009; He et al., 2016b; Yuan et al., 2016; Zhang et al., 2015a; Li et al., 2016)) to filter out users/venues with less than 10 interactions. Table 4.5 summarises the statistics of the filtered datasets.

For each dataset, we measure the quality of the top-K venue recommendations in terms of Mean Average Precision (MAP), Normalised Discounted Cumulative Gain (NDCG) and Mean Reciprocal Rank (MRR), which are widely used for venue recommendation in the literature (Yuan et al., 2016; Wang et al., 2016; Zhao et al., 2014; Loni et al., 2016). These ranking-based metrics were previously described in Section 2.1.2.2. In particular, given the ground truth venues of each user, MAP and MRR consider the ranking nature of the task, by taking into account the rank(s) of the ground truth venues in the produced rankings, while NDCG goes further by considering the checkin frequency/rating value of the user as the graded relevance label. Note that, significance tests are conducted using a paired t-test with $p < 0.01$.

[12] https://snap.stanford.edu/data/

[13] https://www.yelp.com/dataset_challenge

Table 4.5: Statistics of three datasets

	Yelp	Brightkite	Gowalla
Number of users	40,228	25,063	72,953
Number of venues	34,932	48,177	131,328
Number of ratings or checkins	987,050	3,309,555	3,487,258
Number of social links	1,598,096	33,290	330,762
% density of User-Venue matrix	0.0702	0.2740	0.0363

4.3.5 Baselines and Parameter Setup

In this section, we summarise all the baselines used in the experiments to compare the effectiveness of our proposed PRFMC frameworks and its two components: namely the Multi-centre Gaussian and Social power-Law distribution models. Then, we describe the parameters setup for the used baselines in details. We use subscripts to indicate which component is deployed within the PRFMC framework. For example, PRFMC$_M$ and PRFMC$_S$ indicate the PRFMC framework that incorporates the **M**ulti-centre Gaussian model and the **S**ocial power-Law distribution model, respectively. PRFMC$_{MS}$ denotes that the two components of PRFMC are deployed. We compare the effectiveness of each component (i.e. PRFMC$_M$ incorporates the users' geographical movements and PRFMC$_S$ incorporates the social influences) with state-of-the-art venue recommendation approaches that incorporate similar additional sources of information. PRFMC$_{MS}$ incorporates both the users' geographical movements and social influences. In particular, we compare PRFMC$_{MS}$, with a number of baselines, which can be grouped into four categories, namely: traditional BPR, geo-based approaches, social-based approaches and hybrid approaches combining social- and geo-based BPR. All baselines and our proposed PRFMC framework are implemented using LibRec (Guo et al., 2015a), a Java library for recommendation systems. In the following, we summarise our re-implementation of each baseline in details.

4.3.5.1 Traditional BPR

BPR (Rendle et al., 2009) is the classical pairwise ranking approach, coupled with matrix factorisation for user-venue rating/checkin frequency prediction (see Section 2.1.1.1.3 for further details).

4.3.5.2 Geo-based approaches

MGM (Cheng et al., 2012) is a Multi-centre Gaussian Model that incorporates the users' geographical movements. Recommendations are generated by ranking all venues according to the scores computed by Equation (4.19) (see Section 4.3.2 for further details).

GBPR (Yuan et al., 2016) is a state-of-the-art BPR model that leverages the geographical information of venues to sample negative instances. The ranking function of the GBPR model assumes that neighbourhood venues of venues previously visited by users should be ranked higher than the distant ones (see Section 3.5.1 for further details).

4.3.5.3 Social-based approaches

SPLD (Zhang and Chow, 2015) is a Social Power-Law Distribution model that incorporates social influences. In particular, venue recommendations are generated by ranking all venues according to the scores computed by Equation (4.23) (see Section 4.3.3 for further details).

SBPR (Zhao et al., 2014) is a Social BPR model that leverages the users' social links to sample negative instances. The ranking function of the SBPR model assumes that venues previously visited by the user's friends should be ranked higher than venues neither the user nor his/her friends have visited (see Section 3.5.2 for further details).

SWBPR (Wang et al., 2016) is a state-of-the-art BPR model that is extended from SBPR. This model considers the *Strong*- and *Weak*- Social ties of the user's friends. The ranking function of the SWBPR model assumes that venues visited by weak-tie friends should be ranked higher than venues visited by strong-tie friends, because weak tie friends are likely to introduce novel and diverse venues (see Section 3.5.2 for further details).

4.3.5.4 Hybrid (social & geo)-based approaches

GeoSo (Zhang and Chow, 2015) is a state-of-the-art probabilistic model that incorporates both the users' geographical movements and social influences. To permit a fair evaluation, we have re-implemented their GeoSoCa approach to consider only geographical and social information, and ignore the categorical properties of venues, in common with our proposed approach that also does not consider categories. Essentially, the probabilistic scores generated by the GeoSo model is the product of $P_s(i|u)$ in Equation (4.23) and $P_m(i|\mathbb{C}_u)$ in Equation (4.19).

GSBPR is our proposed baseline that combines **GBPR** and **SBPR** together by assuming that the neighbourhood venues visited by the user's friends should be ranked higher than the distant ones. The optimisation criterion of this model is $BPR_{Opt}(D_{gs})$, where:

$$D_{gs} = \left\{(u, i, k, j) \mid i \in V_u^+ \wedge k \in V_{u,i}^g \cap V_{F_u}^s \wedge j \in V_u^-\right\}.$$

Indeed, D_{gs} contains tuples (u, i, k, j) where user u has visited venue i, k is neighbouring venue of venue i that the user has not visited but his/her friends have visited, and j is a venue never visited by neither user u nor by his/her friends.

BPRMC (Loni et al., 2016) is a state-of-the-art BPR model that can simultaneously incorporate multiple sampling approaches (i.e. GBPR and SBPR) based on a pre-defined weight for each sampling approach. This approach is a suitable baseline, as it permits a fair comparison of our proposed PRFMC framework with another model that considers multiple sampling approaches.

To permit a fair comparison, the PRFMC framework and all of the BPR-based baselines deploy the traditional Matrix Factorisation (MF) as the rating/checkin prediction function. Following common practice (Ma et al., 2011; Yuan et al., 2016; Wang et al., 2016; Zhao et al., 2014; Koren et al., 2009), the hyperparameters of the traditional MF model are set as follows: the dimension of the latent factors $d = 10$, and the regularisation parameters $\lambda_u = 0.001, \lambda_v = 0.001$. To the fullest extent possible, we apply the parameters used by the baselines and the probabilistic models (MGM, SPLD and GeoSo) when these were applicable, i.e. when the values reported in the corresponding papers were recommended for the datasets we use in our experiments. For instance, following Cheng et al. (2012), we set the hyperparameters of MGM as follows: $\phi = 0.02$, the distance threshold $\kappa = 15$ and the frequency control parameter $\alpha = 0.2$. The parameters of the SBPR, SWBPR, GBPR models are determined using the validation set for each fold. Similarly, for other approaches not previously reported on these datasets, we determine the values for their parameters using the validation set for each fold.

4.3.6 Experimental Results & Discussion

In this section, to answer research questions **RQ4.3-4.5** described in Section 4.3.4, we evaluate the effectiveness of the PRFMC framework (PRFMC$_{MS}$) and its components (PRFMC$_{M}$ and PRFMC$_{S}$) in comparison with various baselines. We present the results of our experiments in Table 4.6, which reports the effectiveness of various approaches in term of the MAP, NDCG and MRR measures on the three used datasets. The grouped columns of the table correspond to the grouping of baseline approaches based upon the sources of additional information, as discussed in Section 4.3.5, along with the corresponding implementation of PRFMC. To further report the effectiveness of PRFMC in comparison with the baselines, Table 4.7 reports the mean percentage differences across the MAP, NDCG and MRR measures of various approaches compared to BPR on the three different datasets. On inspection of our re-implementations of the baselines in Table 4.6, we note that the relative top-K venue recommendation quality of the baselines on the three datasets in terms of the three effectiveness measures is consistent with the results reported for the various baselines in the corresponding literature (Loni et al., 2016; Wang et al., 2016; Yuan et al., 2016; Zhao et al., 2014). For instance, GBPR outperforms BPR by 3-9% across the three datasets (Yuan et al., 2016) and SWBPR outperforms SBPR by 0.22-25% across the three datasets. Note that previous works (Loni et al., 2016; Wang et al., 2016; Zhao et al., 2014) used different datasets, while our re-implementations of their proposed approaches obtain relatively similar improvements. We now analyse each group of approaches in turn based upon the source of additional information employed.

To answer research question **RQ4.3**, within the geo-based group of columns in Table 4.6, we compare PRFMC$_M$ with the MGM and GBPR models, which correspond to the probabilistic baseline and the extended geo-based BPR baseline that leverage the geographical information of venues to sample negative instances, respectively. Note that the GBPR model does not take the users' geographical movements previously observed in the literature into account, while PRFMC$_M$ and MGM do. First, we observe that the MGM model that leverages the users' geographical movements to generate venue recommendations is less effective than the traditional BPR model in recommending ranked list of venues to the users across all three datasets. This observation is consistent with the literature (Rendle et al., 2009) and is expected as the MGM is a pointwise-based approach, which is likely to be less effective than pairwise-based approaches (e.g. the traditional BPR model) (Rendle et al., 2010). Furthermore, note that the BPR model is only trained on the user-venue interactions and ignores the geographical information of venues. These results demonstrate that the probabilistic model, MGM (pointwise), is largely outperformed by the BPR model (pairwise), which is consistent with the results reported in previous works (Rendle et al., 2009; Liu, 2009).

Next, we observe that PRFMC$_M$ consistently and significantly outperforms the MGM and GBPR models for MAP, NDCG and MRR across all datasets. This implies that our proposed PRFMC framework, which leverages the users' geographical movements captured by the Multi-centre Gaussian model (MGM) to effectively sample negative instances, is more effective than the GBPR model (Yuan et al., 2016), which itself relies on a pre-defined assumption on the likely relevance of neighbouring venues during the negative sampling process, as summarised by **Limitation S2** in Section 3.5.3. In addition, within the geo-based group of columns in Table 4.7, PRFMC$_M$ can enhance the effectiveness of the BPR model by 8%, 26% and 21% on the Yelp, Gowalla and Brightkite datasets, respectively. In contrast, the GBPR model can only improve the performance of the BPR model by 6%, 5% and 4% on the Yelp, Gowalla and Brightkite datasets, respectively. The high margin of improvements

Table 4.6: Performances in terms of MAP, NDCG and MRR of various approaches. For each type of additional information and evaluation measure, the best performing result is highlighted in bold and $*$ indicates significant differences in terms of paired t-test with $p < 0.01$, comparing to the best performing result. Percentage differences compared to BPR are denoted by Δ.

Dataset	Measure		Geo-based			Social-based				Hybrid geo- & social-based			
		BPR	MGM	GBPR	PRFMC$_M$	SPLD	SBPR	SWBPR	PRFMC$_S$	GeoSo	GSBPR	BPRMC	PRFMC$_{MS}$
Yelp	MAP	0.1974	0.0080*	0.2037*	**0.2071**	0.0011*	0.2014*	0.2051*	**0.2101**	0.0062*	0.2016*	0.2042*	**0.2109**
	Δ		-95.94%	3.18%	4.89%	-99.44%	2.03%	3.89%	6.45%	-96.86%	2.12%	3.44%	6.81%
	NDCG	0.3253	0.1575*	0.3467*	**0.3587**	0.1246*	0.3451*	0.3521*	**0.3625**	0.1581*	0.3432*	0.3519*	**0.3690**
	Δ		-51.58%	6.58%	10.25%	-61.70%	6.07%	8.21%	11.43%	-51.41%	5.48%	8.15%	13.41%
	MRR	0.2186	0.0197*	0.2343*	**0.2402**	0.0031*	0.2275*	0.2384*	**0.2506**	0.0137*	0.2295*	0.2361*	**0.2492**
	Δ		-90.96%	7.19%	9.91%	-98.59%	4.11%	9.07%	14.68%	-93.73%	5.01%	8.04%	14.01%
Gowalla	MAP	0.0703	0.0511*	0.0724*	**0.0826**	0.0003*	0.0722*	0.0758*	**0.0843**	0.0503*	0.0724*	0.0732*	**0.0933**
	Δ		-27.35%	3.01%	17.50%	-99.59%	2.79%	7.91%	19.99%	-28.41%	3.07%	4.11%	32.76%
	NDCG	0.2485	0.2307*	0.2578*	**0.2929**	0.1054*	0.2669*	0.2678*	**0.2894**	0.2259*	0.2592*	0.2717*	**0.3174**
	Δ		-7.17%	3.71%	17.83%	-57.60%	7.40%	7.76%	16.43%	-9.11%	4.28%	9.32%	27.69%
	MRR	0.0881	0.1142*	0.0951*	**0.1259**	0.0010*	0.0877*	0.1098*	**0.1364**	0.0779*	0.0958*	0.0906*	**0.1510**
	Δ		29.61%	7.93%	43.00%	-98.85%	-0.42%	24.63%	54.82%	-11.53%	8.78%	2.87%	71.47%
Brightkite	MAP	0.1561	0.0459*	0.1607*	**0.1854***	0.0749*	0.1518*	0.1528*	**0.1649**	0.1132*	0.1470*	0.1525*	**0.1857**
	Δ		-70.62%	2.94%	18.76%	-52.05%	-2.77%	-2.11%	5.62%	-27.46%	-5.81%	-2.34%	18.95%
	NDCG	0.3026	0.1997*	0.3109*	**0.3618**	0.2124*	0.3042*	0.3007*	**0.3165**	0.2816*	0.2939*	0.3055*	**0.3647**
	Δ		-34.00%	2.73%	19.56%	-29.81%	0.53%	-0.62%	4.58%	-6.95%	-2.88%	0.96%	20.50%
	MRR	0.1738	0.0786*	0.1841*	**0.2170**	0.1202*	0.1688*	0.1692*	**0.1916**	0.1577*	0.1614*	0.1696*	**0.2193**
	Δ		-54.77%	5.88%	24.86%	-30.84%	-2.89%	-2.67%	10.24%	-9.25%	-7.14%	-2.41%	26.16%

Table 4.7: Mean percentage differences across MAP, NDCG and MRR of various approaches compared to BPR.

Dataset	Geo-based			Social-based				Hybrid geo- & social-based			
	MGM	GBPR	PRFMC$_M$	SPLD	SBPR	SWBPR	PRFMC$_S$	GeoSo	GSBPR	BPRMC$_{GS}$	PRFMC$_{MS}$
Yelp	-79%	6%	**8%**	-87%	4%	7%	**11%**	-80%	4%	7%	**11%**
Gowalla	-2%	5%	**26%**	-85%	3%	13%	**30%**	-16%	5%	5%	**44%**
Brightkite	-53%	4%	**21%**	-38%	-2%	-2%	**7%**	-15%	-5%	-1%	**22%**

over the BPR model that are achieved by our proposed PRFMC$_M$ framework demonstrate the usefulness of the users' geographical movements, as previously observed in the literature. Therefore, in response to research question **RQ4.3**, we conclude that the users' geographical movements captured by the MGM model can 1) be leveraged to effectively sample negative instances and 2) improve the quality of venue recommendations of the traditional BPR model.

To answer research question **RQ4.4**, we consider the social-based column group, to compare the effectiveness of the PRFMC$_S$ framework with the SPLD, SBPR and SWBPR models. Trends that are similar in nature to those observed for the geo-based approach group are observed. First, the probabilistic SPLD model that leverages the social influences to generate the venue recommendations is outperformed by the traditional BPR model in recommending ranked lists of venues to the users on the three datasets in terms of three measures. Again, these results are consistent with previous works (Rendle et al., 2009; Liu, 2009) where the pointwise model (i.e. SPLD) is observed to be less effective than the pairwise model (i.e. BPR) in generating high quality venue recommendations. Similar to PRFMC$_M$, we observe that PRFMC$_S$ can consistently and significantly outperform the extended social-based BPR baselines that leverage the users' social links (i.e. SBPR and SWBPR models) to sample negative instances based on the pre-defined sampling assumption that venues previously visited by friends are likely to be visited, as summarised by **Limitation S2** in Section 3.5.3.

Interestingly, the relatively low results for SBPR and SWBPR across MAP, NDCG and MRR on the Brightkite dataset are likely due to the sparsity of the social links between the users in the Brightkite LBSN (see Table 4.5). In contrast, PRFMC$_S$ can improve the effectiveness of BPR, whereas SBPR and SWBPR both do not. Indeed, we find that sampling negative venues using the power-law distribution model is more effective than the pre-defined sampling criteria proposed by Wang et al. (2016); Zhao et al. (2014). Moreover, exploiting the Social Power-Law Distribution (SPLD) model to sample negative venues is more useful to enhance the quality of venue recommendations than simply ranking venues according to the score computed by the SPLD model. In particular, PRFMC$_S$ can enhance the effectiveness of the BPR model by 11%, 30% and 7% on the Yelp, Gowalla and Brightkite datasets, respectively. In contrast, the state-of-the-art SWBPR model that leverages similar additional information as PRFMC$_S$ can only improve the performance of the BPR model by 7%, 13% and -2% on the Yelp, Gowalla and Brightkite datasets, respectively. The large improvements over the BPR model that are achieved by our proposed PRFMC$_S$ framework demonstrate the usefulness of the social influences previously observed in the literature. Therefore, in response to research question **RQ4.4**, we conclude that the social influences captured by the SPLD model can (1) be leveraged to effectively sample negative instances and (2) improve the quality of venue recommendations of traditional BPR model.

With respect to research question **RQ4.5**, we consider the deployment of hybrid models that leverage both geo- and social-based additional sources of information (the Hybrid geo- & social-based group of columns in Table 4.6). We compare our proposed PRFMC$_{MS}$ framework with the GeoSo, BPRMC and GSBPR models. In particular, we compare our proposed framework that is comprised of geographical and social components PRFMC$_{MS}$, with the state-of-the-art probabilistic GeoSo model and the state-of-the-art extended BPR models that can incorporate multiple sampling criteria, namely the GSBPR and BPRMC models. Again, trends that are similar in nature to those observed for the geo- and social-based approach groups are observed. First, the state-of-the-art probabilistic GeoSo model that leverages both the users' geographical movements and social influences to generate the venue recommendations is outperformed by the traditional BPR model on the three used datasets in terms of MAP, NDCG and MRR. This confirms that even the state-of-the-art probabilistic model (pointwise) is less effective than the simple pairwise model (BPR) for venue recommendations. Next, we clearly observe that our proposed PRFMC$_{MS}$ framework can consistently and significantly outperform all the hybrid geo- and social-based BPR baselines, namely GSBPR and BPRMC. In particular, within the Hybrid geo- & social-based group columns in Table 4.7, the PRFMC$_{MS}$ framework can improve the effectiveness of the BPR model by 11%, 44% and 22% on the Yelp, Gowalla and Brightkite datasets, respectively. In contrast, across the three used datasets, the GSBPR and BPRMC can only enhance the performance of the BPR model by approximately 4% and 5%, respectively. The large improvements of the BPR model that are achieved by our proposed PRFMC$_{MS}$ framework demonstrate the usefulness of the user's preference score in Equation (4.11) and the pairwise ranking function in Equation (4.12) of PRFMC$_{MC}$ that leverages both the users' geographical movements and social influences captured by the probabilistic MGM and SPLD models, respectively, to effectively sample negative instances.

Next, we discuss the effectiveness of the GSBPR, BPRMC & PRFMC$_{MS}$ models in comparison with each of their constituent geo- and social-based component baselines. In particular, from Table 4.6, we observe that the results of the GSBPR model are generally not higher than both of its constituents that each consider only one sampling criterion (i.e. the

GBPR and SBPR models). This implies that simply combining the sampling criteria (as done by GSBPR) is not a suitable approach. In contrast, the BPRMC model is more effective than GSBPR at combining multiple sampling criteria. Moreover, by comparing BPRMC with the GBPR and SBPR models, we find that, for all three metrics in the Yelp dataset, BPRMC outperforms the extended BPR models that consider only a single sampling criterion (i.e. GBPR and SBPR). However, for the Gowalla and Brightkite datasets, the effectiveness of BPRMC greatly decreases when one of the constituent sampling criterion is not effective. For instance, regarding the results of the GBPR, SBPR and BPRMC models in terms of MAP and MRR in the Brightkite dataset, we observe that when the performance of the SBPR model decreases, the effectiveness of the BPRMC model also decreases. A similar observation is found for BPRMC in terms of MRR in the Gowalla dataset. These results suggest that BPRMC cannot fully leverage the effectiveness of its combined sampling criteria.

Next, we compare the effectiveness of PRFMC, which considers different sampling criterion (i.e. PRFMC$_M$, PRFMC$_S$ and PRFMC$_{MS}$). The results show that our proposed PRFMC$_{MS}$ framework, which samples negative examples based on both the users' geographical movements and social influences – captured by the Multi-centre Gaussian (MGM) model and the Social Power-Law Distribution (SPLD) model, respectively – significantly outperforms both PRFMC$_S$ and PRFMC$_M$, across all three metrics on all three datasets. Overall, the strong results displayed by PRFMC$_{MS}$ demonstrate the effectiveness of PRFMC in combining different types of sampling criteria. In addition, unlike BPRMC, the effectiveness of PRFMC$_{MS}$ does not decrease if one of the fused sampling criteria is not effective. For example, on the Brightkite dataset, PRFMC$_{MS}$ consistently and significantly outperforms its constituents (PRFMC$_M$ and PRFMC$_S$) across the three metrics, while BPRMC is outperformed by one of its constituents (GBPR). Overall, regarding the generalisation of the PRFMC$_{MS}$ framework, the experimental results on the three used datasets reported in Table 4.6 demonstrate that our proposed PRFMC framework appears to be more generalised than the state-of-the-art hybrid geo- and social BPR model, BPRMC. Finally, in response to research question **RQ4.5**, we find that our PRFMC framework provides a significant benefit across various datasets and measures, compared to various existing state-of-the-art single criterion negative sampling approaches (i.e. GBPR, SBPR, SWBPR models), multiple criteria negative sampling approaches (i.e. GSBPR and BPRMC models) as well as probabilistic models (i.e. MGM, SPLD and GeoSo). Indeed, among the results reported in Table 4.6, all of the highest improvements over the traditional BPR model, for all three measures on all three datasets, are observed for the PRFMC$_{MS}$ hybrid negative sampling approach. Indeed, for the Gowalla dataset, PRFMC$_{MS}$ attains a 71% improvement over the MRR of BPR, as well as 37% and 59% improvements in MRR over the recently proposed SWBPR (Wang et al., 2016) and GBPR (Yuan et al., 2016) approaches (Table 4.6: 0.1098→0.1510; 0.0951→0.1510), respectively.

4.4 Conclusions

In this chapter, we proposed the Social and Textual Regularisation (STReg) technique and the textual MF-based approach (MFw2v) that exploit word embeddings to model the semantic properties of the textual content of comments associated with the users' rating to enhance the effectiveness of the traditional MF model for the user-venue rating prediction. Our proposed STReg technique and our MFw2v approach aim to address the limitations of existing works: namely **Limitation M1** of the state-of-the-art Social Regularisation (SoReg) technique and **Limitations M2-M4** of the state-of-the-art textual MF-based approaches (CMF and JMF models). In Section 4.2.4, we empirically evaluated the effectiveness of STReg and MFw2v on the Yelp dataset, which consists of over 2.2 million ratings from 500k users. Our comprehensive experiments demonstrated the usefulness of our prosed STReg technique in improving the prediction accuracy of the traditional MF model. In addition, the STReg technique can outperform the SoReg technique by 3-4% in terms of the RMSE and MAE metrics. Moreover, our proposed MFw2v model is shown to be more effective than the CMF and JMF models and as effective as the STReg technique, while only requiring the venues' comments as auxiliary information and not taking the users' social links into account.

Moreover, in this chapter, we further proposed the Personalised Ranking Framework with Multiple sampling Criteria (PRFMC) that can leverage multiple types of additional information to enhance the effectiveness of the traditional BPR model for top-K venue recommendations. Our proposed PRFMC framework aims to address **Limitations S1 & S2** of existing extended BPR modes (i.e. SBPR, SWBPR and GBPR models). To alleviate these limitations, the PRFMC framework exploits the state-of-the-art probabilistic models (i.e. multi-centre gaussian and the power-law distribution models) that take the users' geographical movements and social influences into account to effectively sample negative instances. In Section 4.3.6, we empirically evaluated the effectiveness of PRFMC in comparison to various BPR-based models on three large-scale datasets: namely the Yelp, Gowalla and Brightkite LBSNs. Our comprehensive experiments demonstrate the effectiveness of our proposed PRFMC framework, which is markedly superior to the state-of-the-art BPR models. For example, the PRFMC framework can improve the effectiveness of the traditional BPR model by approximately 7-71% in terms of MAP, NDCG and MRR, while the recently proposed GBPR, SWBPR and BPRMC models can only enhance the performance of the BPR model by approximately 3-8%, 4-24% and 1-9%, respectively. Moreover on the Gowalla dataset, PRFMC attains a 37% improvement in MRR over the SWBPR model. Moreover, the improvements of the PRFMC framework are attained without increased computational complexity compared to the BPR baselines.

In our thesis statement (see Section 1.2), we hypothesised that the quality of personalised ranked list of venues can be effectively enhanced by leveraging additional information such as the users' social relationships, the textual content of comments and the geographical information of venues. Based upon our experiments in this chapter, we conclude that our proposed STReg technique, our MFw2v model and our PRFMC framework that all leverage these additional information can generate high quality venue recommendations. Note that there is also other type of additional information that can be leveraged to improve the quality of venue suggestion such as temporal and seasonal information[14]. In addition, as we have previously discussed in Chapter 1, the users' observed checkins usually exhibit sequential properties, which can also be leveraged to capture the users' short-term (*dynamic*) preferences. In the next chapter, we propose a novel Deep Recurrent Collaborative Filtering (DRCF) framework that exploits Recurrent Neural Network (RNN) models to effectively capture the users' *dynamic* preferences from the sequence of their checkins.

[14] Later in Chapter 6 and Chapter 7, we will demonstrate how to leverage the temporal information to improve the quality of venue recommendation.

Chapter 5

Deep Recurrent Collaborative Filtering Framework

5.1 Introduction

In Chapter 3, we argued that the users' recent observed checkins can have more influence on their decisions on the next venues that they will visit than their historical checkins (e.g. 6-month-old checkins). As mentioned in Section 3.6.1, the literature on sequential-based recommendation systems (Yu et al., 2016; Tang et al., 2017; Zhang et al., 2014b; Koren, 2010a; Rendle, 2012; Cheng et al., 2013) has already shown that the sequential properties of the user's interactions (e.g. the sequences of checkins, click and purchases) play an important role in improving the quality of recommendations. For example, consider a user who visited several art museums a few months ago and has recently visited a museum and a restaurant, sequentially. Models that only capture the user's long-term (*static*) preferences will recommend other museums to visit, whereas a model that can capture the user's short-term (*dynamic*) preferences might recommend a nearby bar to visit instead. However, traditional MF-based approaches, described in Section 2.1.1.1.2, can only leverage the historical checkins of the users to model their *static* preferences, and therefore ignore the sequential properties of their checkins. We argue that the quality of venue recommendations can be further improved by leveraging the sequential properties of observed checkins to effectively capture the recent so-called *dynamic* preferences of users.

In Section 3.4, we described the state-of-the-art Neural Matrix Factorisation (NeuMF) framework proposed by He et al. (2017), an extension of traditional Matrix Factorisation that exploits Deep Neural Networks (DNN) to capture the complex structure of user-venue interactions. Indeed, the NeuMF framework aims to address the limitation of traditional

MF-based approaches, as discussed in Section 3.4.1, which rely on the dot product of latent factors of users and venues to estimate the user's preference on the venues. However, we argue that there are four limitations related to the NeuMF framework (**Limitations N1-N4**, as elicited in Section 3.4.5), which need to be addressed in order to effectively capture the complex structure of user-venue interactions from their sequential order of checkins for venue recommendations. **Limitations N1-N4** of the NeuMF framework are summarised again below:

Limitation N1: There is a disadvantage in the NeuMF framework for identifying the top-ranked venues to present to users due to its optimisation.

Limitation N2: There is a disadvantage in the NeuMF framework assuming that the users' preferences are *static* and do not account for the sequential properties of observed feedback.

Limitation N3: There is a disadvantage in the NeuMF framework that ignores the dot product of latent factors, which capture user-venue interactions.

Limitation N4: There is a disadvantage in the NeuMF framework that applies the traditional BPR negative sampling approach, in which the sequential order of observed checkins is ignored by the negative sampling approach.

Although several RNN-based factorisation approaches have been proposed in the literature, as previously discussed in Section 3.6.1, those approaches still rely on the dot product operation and can only capture the *dynamic* preference of the users, which may be not sufficient to capture the complex structure of user-venue interactions from their sequences of checkins (**Limitations R1-R2**, as identified in Section 3.6.3). Moreover, these RNN-based factorisation approaches still rely on the traditional negative sampling approach, described in Section 2.1.1.1.3, Algorithm 2.1, which does not take the sequential order of checkins into account during the training process (**Limitation S3**). Table 5.1 provides the summary of these existing approaches and their corresponding limitations. **Limitations R1-R2** and **Limitation S3** can be summarised below:

Limitation R1: The RNN-based factorisation approaches for which this limitation applies (RNN-MF (Zhang et al., 2014b), described in Section 3.6.1) do not take the users' long-term (*static*) preferences into account.

Limitation R2: The RNN-based factorisation approaches for which this limitation applies (RNN-MF (Zhang et al., 2014b) and DREAM (Yu et al., 2016), described in Section 3.6.1) still rely on the dot product operation to combine the latent factors of users and venues as well as the hidden unit when predicting a user's checkin.

Table 5.1: Summary of existing factorisation approaches and its corresponding limitations

	NeuMF	RNN-MF	DREAM	DRCF
Deep Neural Network	✓	✓	✓	✓
Sequential-based	×	✓	✓	✓
Limitations	N1-N4	R1,R2,S3	R1,R2,S3	-

Limitation S3: The sampling approaches for which this limitation applies do not take the sequential order of checkins into account.

In this chapter, we propose a novel Deep Recurrent Collaborative Filtering framework (DRCF), an extension of the NeuMF framework, which exploits Recurrent Neural Networks (RNN) to capture the users' *dynamic* preferences from their sequences of checkins. The DRCF framework consists of three components: namely (i) a Generalised Recurrent Matrix Factorisation (GRMF) model, (ii) a Multi-Layer Recurrent Perceptron (MLRP) model and (iii) a Recurrent Matrix Factorisation (RMF) model. Within the DRCF framework, we propose novel *dynamic* and *static* geo-based negative sampling approaches that take the sequential properties of checkins and geographical location of venues into account to enhance the effectiveness of the DRCF framework, as well as alleviate the cold-start user problem. In particular, the DRCF framework aims to address **Limitations N1-N4**, while our proposed *dynamic* and *static* geo-based negative sampling approaches aim to address both **Limitations N4 & S3**. Moreover, within the components of the DRCF framework, the GRMF and MLRP models aim to address **Limitation R1** and **Limitation R2** of the existing RNN-based factorisation approaches, respectively. The remainder of this chapter is structured as follows:

- Section 5.2 provides an overview of the DRCF framework with a *pairwise* ranking function and the *dynamic* geo-based negative sampling approach for venue recommendation.
- Section 5.3 describes our proposed *dynamic* and *static* geo-based negative sampling approaches that take the geographical information of venues into account during the negative sampling process. In particular, our proposed *dynamic* geo-based negative sampling approach takes the sequential order of the users' checkins during the negative sampling process, whereas *static* geo-based negative sampling approach does not.
- Section 5.4 provides details of the first component of the DRCF framework, the Generalised Recurrent Matrix Factorisation (GRMF) that exploit the RNN models and the element-wise product operation to capture the *dynamic* preferences of users.
- Section 5.5 provides details of the second component of the DRCF framework, the Multi-Layer Recurrent Perceptron (MLRP), which exploits the Multi-Layer Percep-

tron, the RNN models as well as the concatenation operation to model the *dynamic* preferences of users.

- Section 5.6 provides details of the third, an final, component of the DRCF framework, the Recurrent Matrix Factorisation (RMF), which exploit the dot product operation to model both the *static* and *dynamic* preferences of users.
- Section 5.7 presents our experimental methodology in terms of datasets and measures as well as algorithm parameters.
- In Section 5.8, we empirically evaluate the effectiveness of our proposed DRCF framework in comparison with the state-of-the-art NeuMF framework and the RNN-based factorisation approaches discussed in Chapter 3.
- Section 5.9 provides a summary of this chapter.

5.2 Overview of Deep Recurrent Collaborative Filtering (DRCF) Framework

In this section, we provide an overview of our proposed Deep Recurrent Collaborative Filtering (DRCF) framework, which is illustrated in Figure 5.1. The DRCF framework consists of five layers: namely input, embedding, Recurrent Neural Network (RNN), Neural Collaborative Filtering (CF) and output layers. These layers are connected to each other using blue-dashed and red-dotted lines. The DRCF framework differs from the NeuMF framework in several aspects. First, starting at the bottom of Figure 5.1, at the input layer, we extend the NeuMF framework to leverage the sequential order of checkins of each user u at time t, $s_{u,t}$. Then, in the embedding layers, there are four additional embedding layers that are indicated in green circles in Figure 5.1: these are the Generalised Recurrent Matrix Factorisation (GRMF) and the Multi-Layer Recurrent Perceptron (MLRP) embedding layers that are used in the Recurrent Matrix Factorisation (RMF) model, P_{Gd}, Q_{Gd}, P_{Md} and Q_{Md}. The third layer of the DRCF framework consists of the RNN layers that encapsulate the user's *dynamic* preferences, which are highlighted in pink and purple in Figure 5.1. The Neural CF layers are similar to the NeuMF framework except that we include the RMF layers to discover certain latent structures of user-venue interactions, which will be described in Section 5.6. Finally, the output layer of the DRCF framework is the predicted checkin $\hat{c}_{u,i}$, which is defined as follows:

$$\hat{c}_{u,i} = a_{out}(h_{out}(\phi^{GRMF} \oplus \phi^{MLRP} \oplus \phi^{RMF})) \tag{5.1}$$

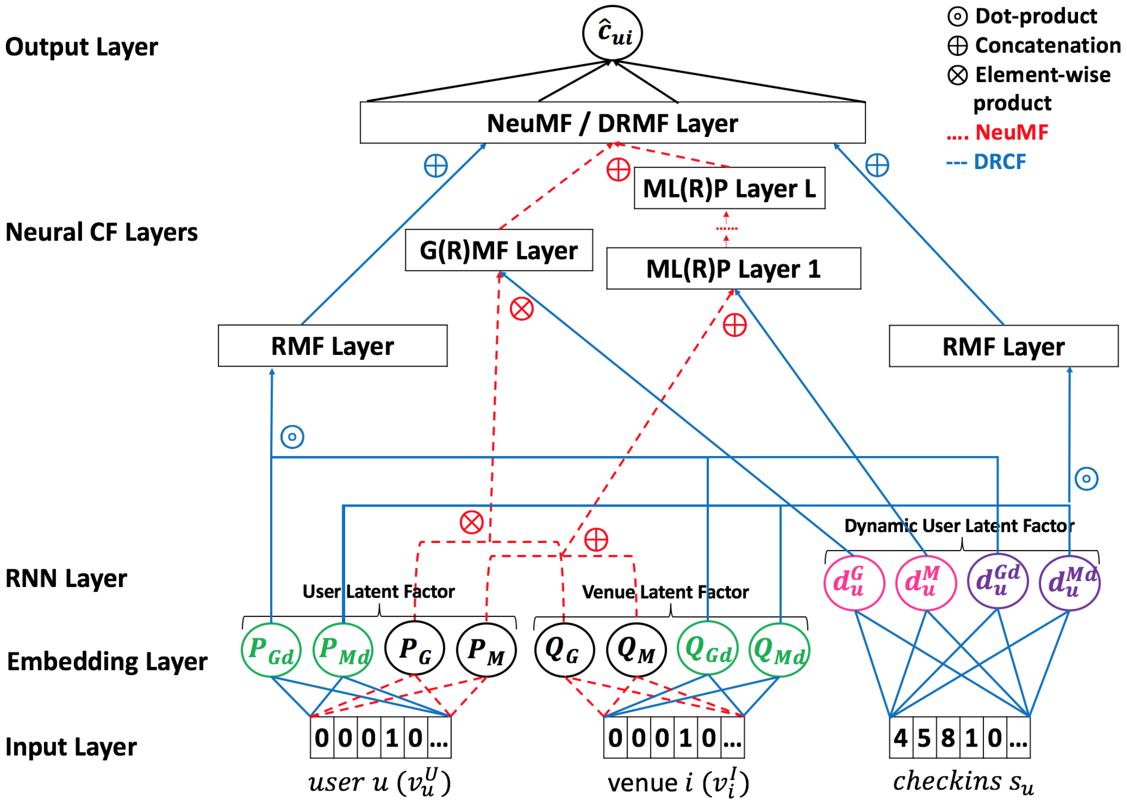


Figure 5.1: Deep Recurrent Collaborative Filtering Framework. The connections of each layer linked by the red-dotted lines illustrate the NeuMF framework. This figure is obtained from Manotumruksa et al. (2017a)

where a_{out} is the activation function, h_{out} is the hidden layer and ϕ^{GRMF}, ϕ^{MLRP} and ϕ^{RMF} denote the GRMF, MLRP and RMF models that will be described in Sections 5.4, 5.5 and 5.6, respectively.

As previously described in Section 3.4.5, the NeuMF framework is trained to minimise the *pointwise* loss between the predicted checkin $\hat{c}_{u,i}$ and the observed checkin $c_{u,i}$, as in Equation (2.4). However, as mentioned in Section 1.1, users in LBSNs only focus on the top-K ranked list of venues, and hence we argue that the training of NeuMF, which aims to minimise a regression metric (e.g. Root Mean Square Error (RMSE)), may not provide an effective top-K ranked list of venues (**Limitation N1**). To address this limitation, we propose to apply Bayesian Personalised Ranking (BPR), described in Section 2.1.1.1.3 (see Equation (2.7)), to learn the parameters $\Theta = \{\theta_e, \theta_r, \theta_h\}$, where $\theta_e, \theta_r, \theta_h$ are the parameters of the embedding, RNN and Neural CF layers of the DRCF framework, respectively, as follows:

$$\mathcal{J}(\Theta) = \sum_{u \in \mathcal{U}} \sum_{s_{u,t} \in S_u} \sum_{i \in s_{u,t}} \sum_{j \in \mathcal{V} - s_{u,t}} \log(\sigma(\hat{c}_{u,i} - \hat{c}_{u,j})) \tag{5.2}$$

where i is the venue most recently visited in $s_{u,t}$, j is an unvisited venue sampled from $\mathcal{V} - s_{u,t}$. Finally, the gradients of $\theta_r, \theta_e, \theta_h$ can be estimated by the back propagation through time algorithm proposed by Rumelhart et al. (1988). In the next section, we describe our proposed sequential geo-based negative sampling approaches that leverage the geographical information of the venues as well as the sequential order of the users' checkins.

5.3 Sequential Geo-based Negative Sampling Approaches

As investigated in Section 4.3.6, various existing negative sampling approaches (i.e. the SBPR[1], SWBPR[2] and GBPR[3] models), which can incorporate additional information, play an important role in enhancing the effectiveness of the traditional BPR modelRendle et al. (2009) (Section 2.1.1.1.3) for top-K venue recommendations. However, as identified in Section 3.4.5, the NeuMF framework still rely on the traditional negative sampling approach, which during the training process. As a consequence, this may degrade the effectiveness of the NeuMF framework (**Limitation N4**). Moreover, previous works (Yu et al., 2016; Zhang et al., 2014b; Tang et al., 2017; Zhu et al., 2017; Beutel et al., 2018; Smirnova and Vasile, 2017) have shown that such sequential properties of checkins can enhance the quality of venue recommendation. However, these existing negative sampling approaches as well as our proposed PRFMC framework described in Section 4.3 do not take the sequential order of users' checkins into account during the sampling process, which can lead to a non-optimal sampling approach (**Limitation S3**, identified in Section 3.5.3).

To address **Limitations S3 & N4**, in this section, we describe our proposed negative sampling approaches that leverages both the geographical information of venues and the sequential order of the users' checkins to effectively sample negative instances to train the DRCF framework. In particular, in contrast to the traditional negative sampling approach of the traditional BPR model (Rendle et al., 2009), which randomly selects negative instances from a *static* pool of negative venues $\mathcal{V}_u^- = \mathcal{V} - \mathcal{V}_u^+$, we propose a novel *dynamic* geo-based negative sampling approach, denoted as DRCF$_{dgeo}$, which can enhance the effectiveness of the DRCF framework and alleviate the cold-start user problem by taking the sequences of checkins $s_{u,t}$ at time t and the geographical location of venue i, i.e. its neighbour venues $\mathcal{N}_i$, into account. In particular, we modify the objective of the DRCF framework (Equation (5.2))

[1] The Social-based Bayesian Personalised Ranking (SBPR) model proposed by Zhao et al. (2014), described in Section 3.5.2

[2] The Strong and Weak social-based Bayesian Personalised Ranking (SWBPR) model proposed by Wang et al. (2016), described in Section 3.5.2

[3] The Geo-based Bayesian Personalised Ranking (GBPR) model proposed by Yuan et al. (2016), described in Section 3.5.1

to incorporate the geographical information of venues and the sequential order of the users' checkins during the sampling process as follows:

$$\mathcal{J}(\Theta) = \sum_{u \in \mathcal{U}} \sum_{s_{u,t} \in S_u} \sum_{i \in s_{u,t}} \sum_{k \in \mathcal{N}_i - s_{u,t}} \sum_{j \in \mathcal{V} - s_{u,t}} \log(\sigma(\hat{c}_{u,i} - \hat{c}_{u,k})) - \log(\sigma(\hat{c}_{u,k} - \hat{c}_{u,j})) \quad (5.3)$$

where i is the most recently visited venue in the sequence of checkins of user u up to time t, $s_{u,t}$. k is an unvisited venue that is nearby to venue i and $\mathcal{N}_i$ is a set of venues that are nearby to venue i. Algorithm 5.1 describes the optimisation algorithm of the DRCF framework as well as the sampling process of our proposed *dynamic* geo-based negative sampling approach. The *dynamic* negative sampling approach (see lines 8-11 in Algorithm 5.1) samples an unvisited neighbouring venue k and an unvisited distance venue j from a *dynamic* pool of negative venues $\mathcal{N}_i - s_{u,t}$ and $\mathcal{V} - s_{u,t}$ not visited by the user in the sequence of checkins $s_{u,t}$ up to time t, respectively, rather than a *static* pool of negative venues as in the traditional negative sampling approach ($\mathcal{V} - \mathcal{V}_u^+$).

Moreover, we also propose a *static* geo-based negative sampling approach, denoted as DRCF$_{sgeo}$, which samples an unvisited neighbouring venue k and a distant negative venue j from a *static* pool of negative venues $\mathcal{N}_i - \mathcal{V}_u^+$ and $\mathcal{V} - s_{u,t}$ not visited by the user in the current sequence of checkins $s_{u,t}$, respectively. Later, in Section 5.8, we evaluate the usefulness of our proposed *dynamic* and *static* geo-based negative sampling approaches (denoted as DRCF$_{dgeo}$ and DRCF$_{sgeo}$, respectively) in enhancing the effectiveness of the DRCF framework as well as alleviating the cold-start problem. In the following sections, we describe the three components of our proposed DRCF frameworks: namely the Generalised Recurrent Matrix Factorisation (Section 5.4), Multi-Layer Recurrent Perceptron (Section 5.5) and Recurrent Matrix Factorisation (Section 5.6).

5.4 Generalised Recurrent Matrix Factorisation (GRMF)

In this section, we describe the first component of the DRCF framework, the Generalised Recurrent Matrix Factorisation (GRMF) model, which is an extension of the GMF model of the NeuMF framework (Section 3.4.3) that exploits Recurrent Neural Network models (RNN) to effectively capture the users' short-term (*dynamic*) and long-term *static* preferences in the collaborative filtering manner. Previous works (Yu et al., 2016; Tang et al., 2017; Zhang et al., 2014b; Koren, 2010a; Rendle, 2012; Cheng et al., 2013) have shown that the users' *dynamic* preferences captured from the sequential order of their *implicit* feedback plays a crucial role in improving the effectiveness of factorisation-based models. For instance, in the evening, users are more likely to visit a bar directly after they have visited a restaurant.

Algorithm 5.1: An Optimisation Algorithm of the DRCF framework

```
Input: $\mathcal{U}$, $\mathcal{V}$, sequences of visited venues $\mathcal{S}$ and neighbour venues of each venue $\mathcal{N}$
Output: $\Theta = \{\theta_r, \theta_e, \theta_h\}$
initial $\theta_r, \theta_e, \theta_h$
$\mathcal{T} \leftarrow 0$ // iteration number
repeat
    for $\mathcal{T} \leftarrow 1$ to $|\mathcal{U}|$ do
        $u \leftarrow$ draw a random user from $\mathcal{U}$
        for $s_t$ in $\mathcal{S}_u$ do
            $i \leftarrow$ a venue most recently visited in $s_{u,t}$
            $k \leftarrow$ draw a random unvisited venue from $\mathcal{N}_i - s_{u,t}$
            $j \leftarrow$ draw a random unvisited venue from $\mathcal{V} - s_{u,t}$
            Compute gradients of $\theta_r, \theta_e, \theta_h$
            Update the above parameters
until convergence;
```

However, such behaviour cannot be captured by the GMF model of the NeuMF framework because it does not take the sequential properties of checkins into account during the training process (**Limitation N2**, identified in Section 3.4.5).

Various RNN-based factorisation approaches (e.g. RNN-MF (Zhang et al., 2014b) and DREAM (Yu et al., 2016), described in Section 3.6.1) have been recently proposed to exploit RNN-based models (e.g. the RNN, LSTM and GRU models, described in Section 2.2.3) to capture the users' *dynamic* preferences. However, these existing RNN-based factorisation approaches still rely on the dot product operation to model the complex structure of user-venue interactions in a Collaborative Filtering manner (**Limitation R2**, identified in Section 3.6.3), which is not effective as identified by He et al. (2017) [4]. Therefore, to address **Limitation N2** of the GMF model of the NeuMF framework and **Limitation R2** of RNN-based factorisation approaches, we extend the GMF model to leverage the user's sequence of checkins $s_{u,t}$, by exploiting an RNN model as follows:

$$\phi^{GRMF} = d_{u,t}^{G} \otimes \phi u_u^G \otimes \phi v_i^G \tag{5.4}$$

where $d_{u,t}^{G}$ is the user's *dynamic* preferences of user u at time t that are projected from the RNN layer. ϕu_u^G are the latent factors of user u that are projected from the GRMF user embedding layer, P_G, shown in a black circle in Figure 5.1. ϕv_i^G are the latent factors of venue i that are projected from the GRMF venue embedding layer, Q_G.

Unlike the RNN-based factorisation approaches described in Section 3.6.1, to capture

[4] More details are described in Section 3.4.1

the complex structure of user-venue interactions in a Collaborative Filtering manner, instead of the dot product operation, our proposed GRMF model relies on the element-wise product $\otimes$ to combine the latent factors of user ϕu_u^G and venue ϕv_i^G as well as the user's *dynamic* preferences $d_{u,t}^G$. Later, in Section 5.8.2, we demonstrate that by addressing **Limitations N2 & R2**, our proposed GRMF model can significantly outperform the GMF model of the NeuMF framework as well as the existing RNN-based factorisation approaches.

5.5 Multi-Layer Recurrent Perceptron (MLRP)

In the previous section, we described our proposed GRMF model, the extension of the GMF model of the NeuMF framework that exploits the RNN models to capture the users' *dynamic* preferences from the users' sequence of checkins. In this section, we describe the second component of the DRCF framework, the Multi-Layer Recurrent Perceptron (MLRP) model, which is an extension of the MLP model of the NeuMF framework that exploits the RNN model to capture the users' *dynamic* and *static* preferences. As previously mentioned in Section 3.4.5, similar to the GMF model, the MLP model is not sufficiently flexible to leverage the sequential order of the users' checkins to capture their *dynamic* preferences (**Limitation N2**). However, we argue that the effectiveness of the MLP model can be enhanced by leveraging the users' sequential properties of checkins.

As mentioned in Section 3.6, various RNN-based factorisation approaches (i.e. the RNN-MF, BiLSTM-MF and DREAM models) have been proposed to leverage the sequential properties of the users to capture their *dynamic* preferences. However, we argue that these existing RNN-based factorisation approaches are not effective in capturing the users' long-term (*static*) preferences (**Limitation R1**). For example, a venue that a user has visited a couple of months ago has less impact on the user's short-term (*dynamic*) preferences than a venue recently visited but still has a large impact on the user's long-term (*static*) preferences. However, given a long sequence of checkins for a user, sophisticated RNN-based models such as the LSTM and GRU are not able to capture the impact of venues the user visited long time ago to model the user's *static* preferences. In particular, LSTM and GRU that include the forget gate are likely to completely ignore the impact of venues the user visited long time ago because these venues have less impact to the user's recent visited venue (Chung et al., 2014; Yu et al., 2016). However, we argue that an accurate model need to effectively capture both the *static* and *dynamic* preference of users in order to generate high quality top-K venue recommendations.

In particular, to address **Limitation N2** of the MLP model of the NeuMF framework and **Limitation R1** of the existing RNN-based factorisation approaches, we propose to extend the MLP model to exploit the RNN model to effectively capture the users' *dynamic* and *static* preferences. In particular, we exploit multiple hidden layers to independently learn the impact of the user's *dynamic* and *static* preferences as follows:

$$\phi^{MLRP} = a_L(h_L(...a_1(h_1(d_{u,t}^M \oplus \phi u_u^M \oplus \phi v_i^M)))) \tag{5.5}$$

where d^M denotes the user's *dynamic* preferences of user u at time t that is projected from the RNN layer, ϕu_u^M and ϕv_i^M are the latent factors of user u and venue i that are projected from the MLRP Embedding layer, respectively (illustrated in Figure 5.1 by the connection between the red-dashed and blue lines under the MLRP layer with the concatenation operation $\oplus$). Note that the latent factors of user u, ϕu_u^M, represent the user's long-term (*static*) preferences, which are similar to the users' latent factors used by the traditional MF-based approaches. Unlike the existing RNN-based factorisation approaches that combine the user's *dynamic* and *static* preferences ($d_{u,t}^M$ and ϕu_u^M, respectively) using the summation operation (see Equation (3.24)), by using the concatenation operation $\oplus$, our proposed MLRP model can treat $d_{u,t}^M$ and ϕu_u^M independently, while the dependencies of $d_{u,t}^M$ and ϕu_u^M are seamlessly captured by hidden layers $h_1..h_L$.

We note that there are differences between ϕu_u^G (ϕv_i^G) ($d_{u,t}^G$) and ϕu_u^M (ϕv_i^M) ($d_{u,t}^M$) in Equations (5.4) & (5.5). Inspired by He et al. (2017), we exploit different embedding and RNN layers for the GRMF and MLRP models in order to independently learn the complex structures of the *dynamic* user-venue interactions from different models. In particular, by having different embedding and RNN layers, the GRMF model can independently capture the interactions using the element-wise product operation, while the MLRP model independently captures the interactions using the concatenation operation. In fact, He et al. (2017) showed that the effectiveness of the NeuMF framework can be improved by allowing different models (i.e. its GMF and MLP models) to learn from different sets of embedding layers. Later, in Section 5.8.2, we show that our proposed GRMF and MLRP models, which use different embedding an RNN layers to independently learn the complex structures of the *dynamic* user-venue interactions, can improve the effectiveness of the DRCF framework. Moreover, we demonstrate that by addressing **Limitation N2** of the MLP model of the NeuMF framework and **Limitation R1** of the existing RNN-based factorisation approaches, our proposed MLRP model can significantly outperform the MLP model and the existing RNN-based factorisation approaches.

5.6 Recurrent Matrix Factorisation (RMF)

Thus far, we have described two components of the DRCF framework: namely the Generalised Recurrent Matrix Factorisation (GRMF) and Multi-Layer Recurrent Perceptron (MLRP) models, which rely on the element-wise product and concatenation operations, respectively, to capture the complex structure of user-venue interactions in a Collaborative Filtering manner. Note that the GRMF and MLRP models are built based on the GMF and MLP models of the NeuMF framework, which rely on the element-wise product and concatenation operations, respectively. However, as argued in Section 3.4.5, the dot product operation may be useful to capture the complex structure of user-venue interactions, which cannot be captured by the element-wise product and concatenation operations (**Limitation N3**).

In this section, we describe the third and last component of the DRCF framework, the Recurrent Matrix Factorisation (RMF) model that relies on the dot product operation to capture the complex structure of user-venue interactions as follows:

$$\phi^{RMF} = \left[(d_{u,t}^{Gd} \otimes \phi u_u^{Gd}) \circledcirc \phi v_i^{Gd}\right] \oplus \left[(d_{u,t}^{Md} \oplus \phi u_u^{Md}) \circledcirc \phi v_i^{Md}\right] \tag{5.6}$$

where $d_{u,t}^{Gd}$ and $d_{u,t}^{Md}$ are the *dynamic* preferences of user u at time t that are projected from the RNN layer. ϕu_u^{Gd} (ϕu_u^{Md}) and ϕv_i^{Gd} (ϕv_i^{Md}) are the latent factors of user u and venue i that are projected from the embedding layer, respectively. Note that, in order to allow the GRMF, MLRP and RMF models to learn independently, we follow He et al. (2017) to train the RMF model by using different sets of embedding and RNN layers (see the green circles in the embedding layer and the purple circles in the RNN layer of Figure 5.1). In doing so, the complex structures of user-venue interactions are independently captured by these three models using different operations (i.e. the GRMF and MLRP and RMF models use the element-wise product, concatenation and dot product operations, respectively). Although our proposed DRCF framework allows different models (i.e. the GRMF, MLRP and RMF models) to learn independently, in the output layer, we concatenate the outputs of the GRMF, MLRP and RMF models and exploit the hidden layer h_{out} to seamlessly and independently integrate the outputs of those models to effectively generate the ranked-list of venues to the users (see Equation (5.1)).

In summary, our proposed DRCF framework, which consists of the GRMF, MLRP and RMF models, can capture the complex structures of user-venue interactions by leveraging the sequential properties of checkins using the dot product, element-wise product and concatenation operations. To the best of our knowledge, the DRCF framework is first to exploit these three operations to capture both the users' *dynamic* and *static* preferences. In the next section, we describe the experimental methodology used to evaluate the effectiveness of the DRCF framework as well as its components in comparison with various baselines.

5.7 Experimental Methodology

In the previous section, we described our proposed Deep Recurrent Collaborative Filtering (DRCF) framework, which consists of three components: namely the Generalised Recurrent Matrix Factorisation (GRMF), Multi-Layer Recurrent Perceptron (MLRP) and Recurrent Matrix Factorisation (RMF) models. We also described our proposed *dynamic* geo-based negative sampling approach that takes the sequential properties of users' checkins as well as the geographical information of venues into account during the sampling process. Our proposed DRCF framework aims to address **Limitations N1-N4** of the state-of-the-art NeuMF framework (see Section 3.4.5) and **Limitations R1-R2** of the existing RNN-based factorisation approaches (see Section 3.6.3), while the *dynamic* geo-based negative sampling approach aims to address **Limitation S3** of the existing negative sampling approaches (see Section 3.5.3). In this section, we describe the experimental methodology used to validate whether our proposed DRCF framework and the *dynamic* negative sampling approach can address **Limitations N1-N4, R1-R2 and S3**. In particular, we aim to answer the following research questions:

RQ5.1 *Can we enhance the effectiveness of the components of the DRCF framework for venue recommendation systems, namely the GRMF and MLRP models, by (a) leveraging the sequential properties of checkins to capture the users'* dynamic *and* static *preferences, (b) incorporating the dot product of latent factors into the models and (c) training those models to generate accurate ranked lists of venues for users?*

RQ5.2 *Are the MF-based models that capture both users'* dynamic *and* static *preferences using either the element-wise product or the concatenation of latent factors more effective than state-of-the-art RNN-based approaches that model both users'* dynamic *and* static *preferences using a dot product of latent factors?*

Furthermore, as discussed in Section 3.5.3, no previous attempt has proposed negative sampling approaches that take both the sequential order of the users' checkins and the geographical location of venues into account to address the cold-start problem. Hence, our third research question:

RQ5.3 *Can our proposed* dynamic *geo-based negative sampling approach that leverages both the sequential properties of checkins and geographical location of venues (a) enhance the effectiveness of DRCF and (b) alleviate the cold-start problem?*

Table 5.2: Summary of each research question and its corresponding success decision and the limitations of the existing approaches.

Research Question	Limitation	Success Decision
RQ5.1	N1-N4	DRCF is more effective than NeuMF.
RQ5.2	R1-R2	GRMF, MLRP and RMF are more effective than GMF and MLP.
RQ5.3	S3	Our proposed *dynamic* geo-based negative sampling approach can enhance the effectiveness of DRCF.

Table 5.2 summarises the research questions we aim to address in this chapter and their corresponding hypothesis. In particular, to demonstrate that our proposed DRCF framework can address **Limitations N1-N4** of the NeuMF framework, we aim to answer research question RQ5.1 by comparing the performances of DRCF and NeuMF. Next, by answering research question RQ5.2, we aim to demonstrate whether the components of the DRCF framework: namely GRMF, MLRP and RMF can address **Limitations R1-R2** of the existing RNN-based factorisation approaches. Finally, by answering research question RQ5.3, we can demonstrate the importance of the sequential order of users' checkins in improving the effectiveness of the negative sampling process.

In the remainder of this section, we describe the experimental setup in terms of datasets and measures (Section 5.7.1), baselines (Section 5.7.2) and algorithm parameters (Section 5.7.3). The experimental results and analysis follow in Section 5.8.

5.7.1 Datasets & Measures

We conduct experiments using publicly available large-scale LBSN datasets. In particular, similar to the experimental setup described in Section 4.3.4, to show the generalisation of our proposed DRCF framework across multiple LBSN platforms and sources of feedback evidence, we use two checkin datasets from Brightkite[5] and Foursquare[6], and a rating dataset from Yelp[7]. Following the common practice for recommendation tasks from the literature (Rendle et al., 2009; He et al., 2016b; Yuan et al., 2016; Zhang et al., 2015a; Li et al., 2016), we remove venues with less than 10 checkins/ratings. The summary of the statistics of the filtered datasets is shown in Table 5.3. We follow previous works (He et al., 2017; Rendle et al., 2009; He et al., 2016b) and adopt a *leave-one-out* evaluation methodology to evaluate the effectiveness of our proposed DRCF framework. In particular, for each user, we

[5] https://snap.stanford.edu/data/

[6] https://archive.org/details/201309_foursquare_dataset_umn

[7] https://www.yelp.com/dataset_challenge

Table 5.3: Statistics of the three used datasets.

	Brightkite	Foursquare	Yelp
Number of normal users	14,374	10,766	38,945
Number of venues	5,050	10,695	34,245
Number of ratings or checkins	681,024	1,336,278	981,379
Number of cold-start users	5,578	154	6903
% density of User-Venue matrix	0.93	1.16	0.07

select her most recent checkin/rating as a ground truth and randomly select 100 venues that she has not checked-in/rated before as the testing set, where the remaining checkins/ratings are used as the training set. The venue recommendation task is thus to rank those 101 venues for each user, aiming to rank highest the recent, ground truth checkin/rating.

We conduct two separate experiments, namely: *Normal Users* (those with 10 checkins or more) and *Cold-start Users* (those with less than 10 checkins) to evaluate the effectiveness of our proposed DRCF framework and its components in the general and cold-start settings. For instance, for the cold-start setting, we only consider users who have *less than* 10 checkins in the testing set during the evaluation. The number of cold-start users for each dataset is shown in Table 5.3. We use the Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) metrics, which are widely used in previous works (He et al., 2017; Yu et al., 2016; Yuan et al., 2016; He et al., 2016b), to measure the quality of the generated top-K venue recommendation. Both HR and NDCG have been previously described in Section 2.1.2.2. In particular, HR considers the ranking nature of the task, by taking into account the rank(s) of the venues that each user has previously visited/rated in the produced ranking, while NDCG goes further by considering the checkin/rating value of the user as the graded relevance label. Lastly, significance tests are conducted using a paired t-test with $p < 0.05$ and $p < 0.01$.

5.7.2 Baselines

In this section, we describe all the baselines we use in comparison with our proposed DRCF framework and its components. These baselines can be categorised into trivial sanity-check baselines, traditional MF-based approaches, RNN-based approaches and Deep Neural Network-based approaches. Note that such approaches may have not be originally proposed for venue recommendation but are sufficiently flexible to do so without any disadvantages. Our baselines are summarised below.

5.7.2.1 Trivial Sanity-Check Baselines

MostPop is a baseline that ranks venues in descending order of the venues' popularities, calculated across all users.

MostVisit is a baseline that ranks venues for a given user in descending order of the venues' popularity for that user.

RecentVisit is a baseline that takes the user's sequential order of checkins into account and recommends the most recently visited venue to the user.

MF (Koren et al., 2009) is the traditional matrix factorisation approach that aims to accurately predict the users' checkin on the unvisited venues.

5.7.2.2 Traditional MF-based approaches

MF (Koren et al., 2009) is the traditional matrix factorisation approach that aims to accurately predict the users' checkin on the unvisited venues (see Section 2.1.1.1.2).

BPR (Rendle et al., 2009) is the classical pairwise ranking approach, coupled with matrix factorisation for user-venue checkin prediction (see Section 2.1.1.1.3).

GBPR (Yuan et al., 2016) is the state-of-the-art BPR model that incorporates geographical influence. This model assumes that the neighbourhood venues of venues previously visited by users should be ranked higher than the distant ones. It uses a *static* negative sampling approach that incorporates the geographical location of venues (see Section 3.5.1).

5.7.2.3 RNN-based approaches

RNN-MF (Zhang et al., 2014b) is the MF-based approach that exploits RNN-based models to predict the user's sequential clicks for sponsored search (see Section 3.6.1).

DREAM (Yu et al., 2016) is the state-of-the-art RNN-based factorisation model that incorporates BPR for ranking optimisation. As DREAM is originally proposed for next shopping-basket recommendation, to permit a fair comparison with our proposed DRCF framework, we reimplement DREAM to treat a single checkin as the shopping-basket purchase (see Section 3.6.1).

5.7.2.4 Deep Neural Network (DNN)-based approaches

NeuMF (He et al., 2017) is the state-of-the-art Neural Matrix Factorisation framework[8], which consists of two components: namely Generalised Matrix Factorisation (GMF) and Multi-Layer Perceptron (MLP) (see Section 3.4.2).

GMF is the component of the NeuMF framework, which models the user-venue interaction using the element-wise product of latent factors (see Section 3.4.3).

MLP is the component of the NeuMF framework that models the user-venue interaction using the concatenation of latent factors (see Section 3.4.4).

We implement all baselines and our proposed approach using Keras[9], a deep learning framework built on top of Theano[10]. Note that, we do not include the Markov Chain-based baselines (e.g. (Rendle et al., 2010)) into our experiments, as experimental results reported by Yu et al. (2016) showed that RNN-based models are more effective than the Markov Chain-based ones. Similarly, we omit state-of-the-art MF-based approaches (e.g. an element-wise Alternating Least Squares (eALS) approach (He et al., 2016b)) since He et al. (2017) showed that the NeuMF framework significantly outperforms such approaches.

5.7.3 Recommendation Parameter Setup

In this section, we describe how we set the parameters of our proposed DRCF framework and the baselines to permit a fair comparison. First, following common practice in previous works (Ma et al., 2011; Yuan et al., 2016; Wang et al., 2016; Zhao et al., 2014), we set the dimensions of the latent factors d of our proposed DRCF framework and all of the MF-based, RNN-based and DNN-based baselines to be identical: $d = 10$ across three datasets [11]. Then, we follow He et al. (2017) and randomly initialise all hidden, embedding and RNN layers' parameters, $\theta_r, \theta_e, \theta_h$, with a Gaussian distribution (with a mean of 0 and standard deviation of 0.01). We apply the mini-batch Adam optimiser (Kingma and Ba, 2014) to optimise those parameters, which yields a faster convergence than the Stochastic Gradient Descent (SGD). In addition, the Adam optimiser automatically adjusts the learning rate for each iteration. We set the learning rate to 0.001[12] and set the batch size to 256. For a fair comparison, the choice of recurrent models for the RNN-based factorisation baselines and our proposed DRCF framework is fixed to the traditional RNN model (Zhang et al., 2014b). Note that

[8] `https://github.com/hexiangnan/neural_collaborative_filtering`

[9] `https://github.com/fchollet/keras`

[10] `http://deeplearning.net/software/theano/`

[11] Later, in Chapters 6 & 7, we revisit this setting.

[12] The default learning rate setting of the Adam optimiser in Keras.

we omit varying the choice of recurrent models (e.g. LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Chung et al., 2014)) and RNN settings, which have already been explored in the literature (e.g. (Tan et al., 2016; Tang et al., 2017)). Finally, to permit a fair comparison between MF-based approaches that exploit a Multi-Layer Perceptron architecture to capture the user-venue interactions using the concatenation of latent factors (i.e. the MLP and MLRP models), we employ three hidden layers, $L = 3$. As the impact of the hidden layer's size L, batch size and dimension size d have been explored in previous work (He et al., 2016b, 2017), we omit varying the size of the hidden layers, batch size and the dimension of the latent factors. Indeed, larger sizes of hidden layers and embedding dimensions may cause over-fitting and degrade the generalisation of the models (He et al., 2016b, 2017).

5.8 Experimental Results

In this section, to answer research questions RQ5.1-RQ5.3 identified in Section 5.7, we evaluate the effectiveness of our proposed DRCF framework and its components: namely the GRMF, MLRP and RMF models in comparison with various baselines described in Section 5.7.2. First, we compare the DRCF framework with the state-of-the-art NeuMF framework and the traditional MF-based and RNN-based baselines under the *Normal Users* and *Cold-Start Users* settings in Section 5.8.1. Then, we investigate the effectiveness of each component of the DRCF framework (i.e. the GRMF and MLRP models) in comparison with each component of the NeuMF framework (i.e. the GMF and MLP models) in Section 5.8.2. Finally, we investigate the usefulness of our proposed *static* and *dynamic* geo-based negative sampling approaches in enhancing the effectiveness of the DRCF framework and alleviating the cold-start user problem is investigated in Section 5.8.3.

5.8.1 The Effectiveness of the DRCF Framework

Table 5.4 reports the effectiveness of our proposed DRCF framework in comparison to the baselines, including the NeuMF framework as well as the MF-based and RNN-based approaches in term of HR@10 and NDCG@10 on the three used datasets for the *Normal Users* setting. Firstly, on inspection of our reimplementations of the baselines in Table 5.4, we note that the relative venue recommendation quality of the baselines on the three used datasets in terms of the two measures are consistent with the results reported for the baselines in the corresponding literature (He et al., 2017; Yu et al., 2016; Zhang et al., 2014b). For instance, the NeuMF framework consistently outperforms the traditional MF and BPR models across the three datasets. These results demonstrate that the element-wise product and concatenation

Table 5.4: Performances in terms of HR@10 and NDCG@10 of various approaches for the Normal Users setting. The best performing result is highlighted in bold; $-$ and $*$ denote a significant difference compared to the best performing result, according to the paired t-test for $p < 0.05$ and $p < 0.01$, respectively.

	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
MostPop	0.1462*	0.1010*	0.2009*	0.1167*	0.0739*	0.0334*
MostVisit	0.4032*	0.3473*	0.4733*	0.4290*	0.1083*	0.0528*
RecentVisit	0.4809*	0.4370*	0.4584*	0.4037*	0.1096*	0.0542*
MF	0.6206*	0.3470*	0.6656*	0.3818*	0.3539*	0.1734*
RNN-MF	0.6368*	0.3824*	0.8040*	0.5459*	0.3814*	0.1891*
BPR	0.6890*	0.4333*	0.7550*	0.4834*	0.4963*	0.2676*
DREAM	0.7041*	0.4839*	0.8147*	0.6081*	0.4349*	0.2235*
NeuMF	0.7073*	0.5358*	0.8361*	0.5842*	0.4934*	0.2729*
GRMF	0.7363	0.5670*	0.8805*	0.6814*	0.5209*	0.2890*
MLRP	0.7291*	0.5790*	0.8873*	0.7046*	0.4771-	0.2652*
DRCF	**0.7419**	**0.6048**	**0.8952**	**0.7223**	**0.5162**	**0.2963**

of the latent factors are more effective than the dot product of the latent factors in capturing the complex structure of the user-venue interactions. Moreover, regarding our reimplementations of the RNN-based factorisation baselines, we observe that the RNN-MF (DREAM) model can consistently outperform the traditional MF (BPR) model across the three datasets. In particular, these results show that the sequential properties of the users' checkins captured by the RNN layer can improve the quality of top-K venue recommendations, which are consistent with results reported in the literature (Yu et al., 2016; Zhang et al., 2014b). It is of particular note that, although the experiments in previous works (He et al., 2017; Yu et al., 2016; Zhang et al., 2014b) were conducted using different datasets, our reimplementations of their proposed approaches and framework obtain similar relative improvements.

Comparing the DRCF framework with the various baselines in Table 5.4, we observe that the DRCF framework consistently and significantly outperforms all the baselines for HR and NDCG, across all three datasets. In particular, DRCF improves NDCG by 12% ($0.5358 \rightarrow 0.6048$), 23% ($0.5848 \rightarrow 0.7223$) and 8% ($0.2729 \rightarrow 0.2963$) over NeuMF for the Brightkite, Foursquare and Yelp datasets, respectively. These results imply that, using the same source of information, our proposed DRCF framework that takes both users' *dynamic* and *static* preferences as well as the dot products of latent factors into account is more effective than the NeuMF framework (He et al., 2017) that considers only the users' *static* preferences. In addition, we observe that the DRCF framework, which combines both the GRMF and MLRP models consistently and significantly outperforms its individual component, GRMF and MLRP, for both measures across the three datasets, except for HR on the Brightkite dataset, where GRMF is statistically indistinguishable from DRCF (difference in

HR $< 1\%$). As mentioned in Section 5.6, although DRCF allows the GRMF and MLRP models to learn independently by using different embedding and RNN layers, as already explained in Section 5.5, these results show that DRCF can seamlessly and dependently integrate the outputs of these models by exploiting the hidden layer h_{out} (Equation (5.1)) to effectively generate the ranked-list of venues to the users. In particular, the hidden layers h_{out} of the DRCF framework learn the importance of the outputs of GRMF and MLRP models.

Comparing with the state-of-the-art RNN-based factorisation baseline, DRCF consistently and significantly outperforms DREAM in term of NDCG by 24.98% $(0.4839 \rightarrow 0.6081)$, 18.78% $(0.2235 \rightarrow 0.6048)$ and 32.57% $(0.7223 \rightarrow 0.2963)$ for the Brightkite, Foursquare and Yelp datasets, respectively. These high improvements of DRCF over DREAM are likely due to the element-wise product and concatenation of the latent factors used in DRCF. In contrast, DREAM only relies on the dot product of the latent factors, which has already been previously shown in previous work (He et al., 2017) not to be effective in capturing the complex structure of the user-venue interactions. Hence, DREAM is less effective than DRCF in capturing the users' *dynamic* and *static* preferences from their sequence of checkins. Next, we note that unlike the Brightkite and Foursquare checkin datasets, the Yelp dataset consists of only user-venue ratings, and hence the sequential properties of visits to venues are less likely to be observed. We observe that the RNN-based approaches (RNN-MF and DREAM) that consider the users' *dynamic* preferences are more effective than the traditional MF-based approaches (MF and BPR) across the Brightkite and Foursquare checkin datasets, while they are outperformed by BPR on the Yelp dataset since those RNN-based approaches cannot leverage the sequential properties of the Yelp rating data. However, our proposed DRCF framework, which considers both the users' *dynamic* and *static* preferences, is still the most effective across the three types of datasets.

Finally, we demonstrate the effectiveness of the DRCF framework in comparison with the baselines over different iterations at the testing time. Figure 5.2 reports the test performances of DRCF and the baselines on the three used datasets, with respect to the number of iterations. Overall, from the figure, across the three datasets, we observe that DRCF outperforms all of the baselines at every iterations and converges faster than the others. For example, on the Brightkite dataset, DRCF outperforms all of the baselines at every iteration. Moreover, DRCF converges faster than RNN-MF, DREAM and NeuMF, where DRCF converges at 15 iterations, while RNN-MF, DREAM and NeuMF converge after 50 iterations. Similar trends can be observed on the Foursquare dataset, where DRCF consistently outperforms all of the baselines at every iteration in terms of HR and NDCG. Again, DRCF converges faster than all baselines, since DRCF converges at around 25 iterations, while all baselines get converge after 50 iterations. Regarding the performances of various ap-

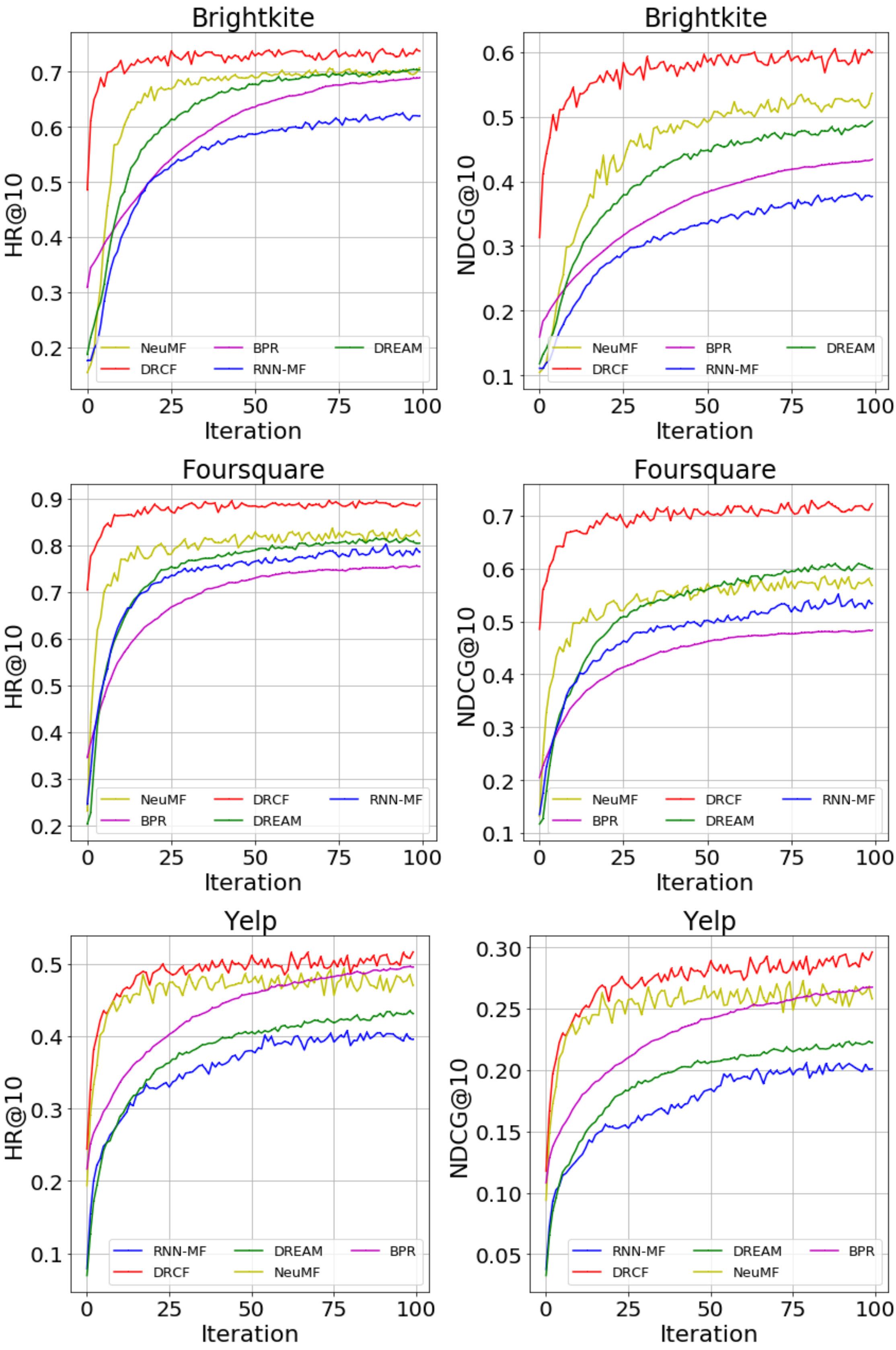


Figure 5.2: Test recommendation HR & NDCG of various approaches with respect to the number of applied iterations.

proaches on the Yelp dataset, where the sequential order of the user-venue rating feedback is less likely to be observed in Figure 5.2, we observe that DRCF can still outperform all the baselines in terms of NDCG at every iterations. Interestingly, BPR and the state-of-the-art NeuMF framework, which can only capture the users' long-term (*static*) preferences, outperform RNN-based factorisation baselines (RNN-MF and DREAM) that can only capture the users' *dynamic* preferences at every iteration. These results are intuitive because the users' short-term (*dynamic*) preferences are less likely to be present on the Yelp dataset. In contrast, our proposed DRCF framework, which can effectively capture both the users' *dynamic* and *static* preferences, outperforms BPR and NeuMF in term of the NDCG metric. DRCF is also as effective as BPR and NeuMF in terms of the HR metric.

Overall, the results from Table 5.4 and Figure 5.2 demonstrate that our proposed DRCF framework is more effective and generalisable than various existing state-of-the-art approaches across the three used datasets in terms of the HR and NDCG measures. In the next section, we investigate the effectiveness of the individual components of the DRCF framework, the GRMF and MLRP models in comparison with the GMF and MLP models of the NeuMF framework, respectively.

5.8.2 The Effectiveness of DRCF's Components

In this section, we further analyse the effectiveness of our proposed DRCF framework by comparing its components (GRMF and MLRP) with the corresponding components of the NeuMF framework (GMF and MLP) as well as the RNN-based approaches (RNN-MF and DREAM). Recall that our proposed GRMF and MLRP models both consist of three components: the RMF layer that incorporates the dot product of latent factors; the RNN layer that models the users' *dynamic* preferences; and BPR for pairwise ranking optimisation (instead of using a pointwise loss function). Hence, to determine the importance of the GRMF's and MLRP's components, we follow an ablation methodology, by recording the effectiveness of the GRMF and MLRP models when each of those three components is removed in turn. For simplicity, we denote d as the RMF layer, r as the RNN layer and b as the BPR optimiser. For example, GRMF$_{rb}$ denotes that the RMF layer is removed from the model, while GRMF$_{rdb}$ denotes that the GRMF model consisting of all three components.

Similar to Table 5.4, Table 5.5 reports the performances of the components of our proposed DRCF framework (GRMF$_{rdb}$ and MLRP$_{rdb}$) in comparison with the components of the NeuMF framework (GMF and MLP), the RNN-based approaches (RNN-MF and DREAM) as well as the corresponding component ablation of GRMF and MLRP. First, we observe similar tends as for the DRCF framework reported in Section 5.8.1. Indeed, we observe that

Table 5.5: Performances in terms of HR@10 and NDCG@10 of various approaches. The best performing result is highlighted in bold; $-$ and $*$ denote a significant difference compared to the best performing result, according to the paired t-test for $p < 0.05$ and $p < 0.01$, respectively.

	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
Component Ablation of GRMF						
RNN-MF	0.6368*	0.3824*	0.8040*	0.5459*	0.3814*	0.1891*
DREAM	0.7041*	0.4839*	0.8147*	0.6081*	0.4349*	0.2235*
GMF	0.7072*	0.4500*	0.7753*	0.4874*	0.4809*	0.2570*
GRMF$_{dr}$	0.7380*	0.5199*	0.8523*	0.6126*	0.4383*	0.2232*
GRMF$_{db}$	**0.7460**	0.5326*	0.8281*	0.5765*	0.5164-	0.2864-
GRMF$_{rb}$	0.6704*	0.4772*	0.8273*	0.5984*	**0.5210**	0.2841*
GRMF$_{rdb}$	0.7363*	**0.5670**	**0.8805**	**0.6814**	0.5209	**0.2890**
Component Ablation of MLRP						
RNN-MF	0.6368*	0.3824*	0.8040*	0.5459*	0.3814*	0.1891*
DREAM	0.7041*	0.4839*	0.8147*	0.6081*	0.4349*	0.2235*
MLP	0.6780*	0.4805*	0.7638*	0.4846*	0.4656*	0.2492*
MLRP$_{dr}$	0.7185*	0.4536*	0.8755*	0.5627*	0.4121*	0.2131*
MLRP$_{db}$	0.6851*	0.4923*	0.8012*	0.5326*	0.4740*	0.2604*
MLRP$_{rb}$	0.6985*	0.5390*	0.8709*	0.6737*	**0.4917**	**0.2705**
MLRP$_{rdb}$	**0.7291**	**0.5790**	**0.8873**	**0.7046**	0.4771*	0.2652*

our proposed GRMF and MLRP models consistently and significantly outperform the RNN-MF, DREAM, GMF and MLP models for the two measures across the three datasets. For example, GRMF (MLRP) improves NDCG by 26% (28%), 39% (45%) and 12% (6%) over GMF (MLP) for the Brightkite, Foursquare and Yelp datasets, respectively. These results show that the models that can capture both users' *dynamic* and *static* preferences (GRMF and MLRP) are more effective that the models that can only capture the users' *static* preferences (GMF and MLP).

Regarding the specific impacts of the three components of GRMF and MLRP (c.f. research question RQ5.1), in general, we observe that all three components play important roles in the effectiveness of the GRMF and MLRP models since significant decreases are often observed when each component is removed. More specifically, regarding the research question RQ5.1(a), with respect to the impact of the sequential properties of the users' feedback as modelled by the r RNN layers, significant decreases for both GRMF$_{rdb}$ and MLRP$_{rdb}$ are consistently observed compared to GRMF$_{db}$ and MLRP$_{db}$, respectively. These results suggest that the users' *dynamic* preferences can significantly improve the effectiveness of the GRMF and MLRP models. In particular, we observe the largest decreases in term of NDCG by approximately 25% in the Brightkite and Foursquare datasets, when the r RNN layers are removed from the MLRP model (see MLRP$_{db}$ in Table 5.8.2). Moreover, as mentioned earlier, since the sequential properties of users' rating feedback on the Yelp dataset are less likely to be observed compared to the Brightkite and Foursquare checkin datasets, we observe that the performances of GRMF$_{rdb}$ and MLRP$_{rdb}$ are significantly decreased by approximately 1.7% in terms of NDCG on the Yelp dataset, when the r RNN layers are ablated.

Next, regarding the research question RQ5.1(b), which concerns the impact of the dot product operation (d) in capturing the complex structure of user-venue interactions, similar trends are observed as in research question RQ5.1(a) where significant decreases for both GRMF$_{rdb}$ and MLRP$_{rdb}$ are consistently observed compared to GRMF$_{rb}$ and MLRP$_{rb}$. In particular, we observe the largest decrease when the RMF layer (d) is removed from the GRMF model for the checkin datasets. For example, the performances of GRMF$_{rdb}$ and MLRP$_{rdb}$ are significantly decreased by 18% and 21% ($0.4772 \rightarrow 0.567$) and ($0.539 \rightarrow 0.579$), respectively, in terms of NDCG on the Brightkite dataset when the RMF layer is ablated. In response to research question RQ5.1(b), these results show that the RMF layer plays an important role in both the GRMF and MLRP models and can significantly improve the effectiveness of the GRMF and MLRP models.

Regarding the research question RQ5.1(c), concerning the impact of the BPR optimiser (b) in generating the top-K venue recommendation, similar trends are observed as in research questions RQ5.1(a) and RQ5.1(b) where significant decreases for both GRMF$_{rdb}$ and MLRP$_{rdb}$ are consistently observed across the three used datasets when these two models are degraded into *pointwise*-based approaches (c.f. GRMF$_{dr}$ and MLRP$_{dr}$). These results demonstrate that the BPR optimiser plays an important role in enhancing the quality of top-K venue recommendation in both the GRMF and MLRP models.

Finally, to answer research question RQ5.2, we compare the effectiveness of GRMF$_{rb}$ and MLRP$_{rb}$ with a state-of-the-art *pairwise* sequential-based approach (i.e. DREAM). First, comparing GRMF$_{rb}$ and DREAM, we observe that GRMF$_{rb}$ can outperform DREAM on the Yelp dataset in terms of both metrics, while GRMF$_{rb}$ is as effective as DREAM on the Foursquare dataset. However, among the three models, MLRP$_{rb}$ is the most effective model, as it can effectively capture both the users' *dynamic* and *static* preferences from the users' sequence of observed feedback. Unlike the GRMF and DREAM models, as described in Section 5.5, by using the concatenation operation, the MLRP model treats the users' *dynamic* and *static* preferences independently, while the dependencies between the *dynamic* and *static* preferences are seamlessly captured by hidden layers (see Equation (5.5)). In response to research question RQ5.2, we find that the concatenation is the most effective operation to effectively combine both users' *dynamic* and *static* preferences, compared to the element-

Table 5.6: Performances in terms of HR@10 and NDCG@10 between various approaches. The best performing result is highlighted in bold; $-$ and $*$ denote a significant difference compared to the best performing result, according to the paired t-test for $p < 0.05$ and $p < 0.01$, respectively.

	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
DRCF	0.7419*	0.6048*	0.8952*	**0.7223**	0.5162*	0.2963*
GBPR	0.7339*	0.4672*	0.8216*	0.5395-	0.5570*	0.3032*
DRCF$_{sgeo}$	0.7847	0.6047*	0.9086	0.7217	**0.5682**	**0.3134**
DRCF$_{dgeo}$	**0.7852**	**0.6210**	**0.9095**	0.7214	0.5618*	0.3064*

wise and dot product operations.

Overall, the strong results from Table 5.5 for the GRMF and MLRP models demonstrate that all the three components (i.e. the (r) RNN layer, the (d) RMF layer and the (b) BPR optimiser) play important roles in the effectiveness of GRMF and MLRP models as significant decreases are often observed when each component is ablated. Moreover, among the three models that can capture both the users' *dynamic* and *static* preferences, MLRP is the most effective. In the next section, we investigate the usefulness of our proposed *dynamic* geo-based negative sampling approaches described in Section 5.3, in enhancing the effectiveness of DRCF in comparison with the state-of-the-art negative sampling approach (GBPR).

5.8.3 The Usefulness of Sequential-based Negative Sampling Approaches

Table 5.6 reports the improvements of the DRCF framework when incorporating our proposed *dynamic* and *static* geo-based negative sampling approaches that take the geographical location of venues into account. In particular, DRCF$_{dgeo}$ denotes the DRCF framework that incorporates our proposed *dynamic* geo-based negative sampling approach, while DRCF$_{sgeo}$ denotes the DRCF framework that incorporate our proposed *static* geo-based negative sampling approach (see Equation (5.3) in Section 5.3). Note that DRCF denotes the DRCF framework that uses the traditional BPR negative sampling approach (Equation (5.2). First, we observe that our proposed *dynamic* negative sampling approach, DRCF$_{dgeo}$, can significantly improve the effectiveness of DRCF in terms of HR by 5.8% and 1.6% on the Brightkite and Foursquare datasets, respectively. In contrast, for the Yelp dataset, where the sequential properties of the users' rating feedback are less likely to be observed, our proposed *static* negative sampling approach, DRCF$_{sgeo}$ can significantly improve the effectiveness of DRCF for the Yelp dataset by 8.8% and 3.4% in terms of HR and NDCG, respectively. Moreover,

Table 5.7: Performances in terms of HR@10 and NDCG@10 of various approaches for Cold-Start Users setting. The best performing result is highlighted in bold; $-$ and $*$ denote a significant difference compared to the best performing result, according to the paired t-test for $p < 0.05$ and $p < 0.01$, respectively.

	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
DRCF vs. Baselines						
MostPop	0.1155*	0.0778*	0.0584*	0.0286*	0.0714*	0.0316*
MostVisit	0.4285*	0.3789*	0.3506*	0.3175*	0.1044*	0.0489*
RecentVisit	0.4995*	0.4585*	0.3831*	0.3446*	0.1052*	0.0497*
MF	0.6768*	0.3913*	0.6623*	0.3650*	0.3748*	0.1868*
BPR	0.7519	0.4907*	0.7792-	0.4961*	0.5273-	0.2946*
RNN-MF	0.6486*	0.3694	0.5909	0.4041*	0.3856*	0.1901*
DREAM	0.7452	0.4969*	0.7987	0.5379*	0.4523*	0.2239*
NeuMF	0.7160*	0.5894-	0.7922	0.6227-	0.5102*	0.2734*
GRMF$_{rdb}$	0.7409-	0.5618*	**0.8442**	0.6542	**0.5399**	0.3083
MLRP$_{rdb}$	0.7418-	0.5779*	0.8377	0.6138*	0.4928*	0.2788*
DRCF	**0.7526**	**0.5980**	0.8377	**0.6645**	0.5330	**0.3136**
Dynamic and *Static* Geo-based Negative Sampling						
DRCF	0.7526*	0.5980*	0.8377-	0.6645-	0.5330*	0.3136*
GBPR	0.8093	0.5262*	0.7468-	0.4717*	0.5802-	0.3202*
DRCF$_{sgeo}$	0.8041	0.6009*	0.8636	0.6748-	**0.5948**	**0.3410**
DRCF$_{dgeo}$	**0.8094**	**0.6199**	**0.8896**	**0.7074**	0.5877	0.3318-

the *static* geo-based sampling approach (DRCF$_{sgeo}$) significantly outperforms the *dynamic* sampling approach (DRCF$_{dgeo}$) in terms of both metrics. Note that there is no significant differences between DRCF, DRCF$_{sgeo}$ and DRCF$_{dgeo}$ in terms of NDCG for the Foursquare dataset. Next, comparing with the state-of-the-art geo-based negative sampling approach GBPR, we find that both DRCF$_{sgeo}$ and DRCF$_{dgeo}$ can significantly and consistently outperform GBPR in terms of the HR an NDCG metrics across the three used datasets, except on Brightkite in terms of HR, where there is no significant difference between GBPR and DRCF$_{dgeo}$. Overall, in response to research question RQ5.3(a), our proposed *dynamic* and *static* geo-based sampling approaches can significantly improve the effectiveness of DRCF across three datasets and are more effective than the state-of-the-art geo-based negative sampling approach (GBPR). In the next section, we evaluate the effectiveness of our proposed DRCF framework and the *dynamic* and *static* geo-based negative sampling approaches in alleviating the cold-start problem.

5.8.4 Cold-Start Users Experiments

In this section, we evaluate the effectiveness of our proposed DRCF framework as well as our proposed negative sampling approaches in alleviating the cold-start user problem. Table 5.7 reports the effectiveness of various approaches in terms of the HR@10 and NDCG@10 on the three used datasets for Cold-Start users. The table contains two groups of rows. The first group reports the effectiveness of the DRCF framework in alleviating the cold-start problem in comparison with the NeuMF framework as well as RNN-based factorisation approaches. The second group reports the improvement of DRCF framework that incorporates our proposed sequential-based negative sampling approaches, which take the geographical location of venues into account (i.e. DRCF$_{sgeo}$ and DRCF$_{dgeo}$).

First, within the first group of rows in Table 5.7, we observe that DRCF can consistently and significantly outperform all the baselines across the three used datasets in terms of NDCG. In particular, DRCF improves NDCG by approximately 1% ($0.5894 \rightarrow 0.5980$), 6% ($0.6227 \rightarrow 0.6645$) and 14% ($0.2734 \rightarrow 0.3136$) over the NeuMF for the Brightkite, Foursquare and Yelp datasets, respectively. Comparing with the state-of-the-art RNN-based factorisation approach, DRCF improves NDCG by 20% ($0.4969 \rightarrow 0.598$), 23% ($0.5379 \rightarrow 0.6645$) and 40% ($0.2239 \rightarrow 0.3136$) over DREAM on the Brightkite, Foursquare and Yelp datasets, respectively. Note that, although GRMF is the most effective model in alleviating the cold-start problem in terms of HR metric on the Foursquare and Yelp datasets, there is no significant difference between GRMF and DRCF. These results suggest that, using the same source of information, our proposed DRCF framework is more effective than the state-of-the-art NeuMF and DREAM in generating venue recommendations for the cold-start users.

Next, regarding research question RQ5.3(b), we investigate the usefulness of our proposed sequential geo-based negative sampling approaches (i.e. DRCF$_{sgeo}$ and DRCF$_{dgeo}$) in alleviating the cold-start problem. Within the second group of rows in Table 5.7, by incorporating either our proposed *dynamic* or *static* geo-based sampling approach into the DRCF framework, DRCF can be significantly improved for both measures, across all three datasets. In particular, for the Brightkite and Foursquare datasets, DRCF$_{dgeo}$ improves NDCG by approximately 3.5% and 6.5% over DRCF, respectively. Moreover, we observe similar trends to those previously observed in Table 5.6. For example, DRCF$_{sgeo}$ is more effective than DRCF$_{dgeo}$ on the Yelp dataset, where the sequential properties of the users' rating feedback are less likely to be observed. In addition, DRCF$_{sgeo}$ can improve both HR and NDCG by approximately 10% and 5.8% over DRCF, respectively, on the Yelp dataset under the Cold-Start Users setting. Therefore, in response to research question RQ5.3(b), we find that our proposed *dynamic* and *static* geo-based sampling approaches that take the geographical information of venues into account can effectively alleviate the cold-start problem of the DRCF framework.

5.9 Conclusions

In this chapter, we proposed a novel Deep Recurrent Collaborative Filtering (DRCF) framework with a pairwise ranking function for venue recommendation. The DRCF framework is an extension of the NeuMF framework that exploits Recurrent Neural Networks to effectively capture the users' *dynamic* preferences from their sequences of checkins. In particular, the DRCF framework consists of three main components: namely the Generalised Recurrent Matrix Factorisation (GRMF), Multi-Layer Recurrent Perceptron (MLRP) and Recurrent Matrix Factorisation (RMF) models. By combining these three models, we showed that the DRCF framework can effectively capture the complex structure of user-venue interactions based on the element-wise product, dot product as well as the concatenation of the latent factors. In addition, within the DRCF framework, we proposed two novel *dynamic* and *static* geo-based negative sampling approaches that take the users' sequential properties of checkins and the geographical information of venues into account to further enhance the effectiveness of the DRCF framework, as well as alleviate the cold-start user problem.

Our proposed DRCF framework and it components aim to address the seven elicited limitations of the state-of-the-art approaches: namely **Limitations N1-N4** of the NeuMF framework and **Limitations R1-R2** of the RNN-based factorisation approaches. Our proposed negative sampling approaches aim to address **Limitation S3** of the existing negative sampling approaches. Our comprehensive experiments on three large-scale datasets from the Brightkite, Foursquare and Yelp LBSNs demonstrate the significant improvements of our proposed DRCF framework and its components as well as our proposed sequential geo-based sampling approaches for venue recommendation in comparison with various state-of-the-art venue recommendation approaches in both normal and cold-start settings. In particular, regarding the research question RQ5.1 (see Table 5.2), our experimental results demonstrate that our proposed DRCF framework can effectively address **Limitations N1-N4** of the state-of-the-art NeuMF framework where DRCF can consistently and significantly outperform NeuMF across the three used datasets in terms of the HR and NDCG metrics by approximately 8-23%. In response to research question RQ5.2, our ablation experiments showed that each component of the DRCF framework: namely GRMF and MLRP can address **Limitations R1 & R2** of the state-of-the-art RNN-based factorisation approach (DREAM), as GRMF and MLRP can significantly and consistently outperform DREAM across different datasets. In addition, in response to research question RQ5.3, the experimental results in Section 5.8.3 demonstrated that our proposed sequential geo-based negative sampling approaches can address **Limitation S3** of the existing negative sampling approaches by comparing the effectiveness of DRCF with or without our proposed *dynamic* and *static* negative sampling approaches and the state-of-the-art geo-based negative sampling approach (GeoBPR). In particular, the obtained experimental results in Section 5.8.3 showed that our proposed *dynamic* negative sampling approach can further enhance the effectiveness of DRCF for both normal and cold-start settings by approximately 3-10% in term of NDCG.

In our thesis statement (see Section 1.2), we hypothesised that the quality of the personalised ranked list of venues can be effectively enhanced by leveraging the sequential properties of the users' checkins. To achieve this, an effective Collaborative Filtering-based framework that models the user's long- (*static*) and short-term (*dynamic*) preferences from the sequence of user's checkins is needed. Based upon our experiments in this chapter, we conclude that our proposed DRCF framework as well as our two proposed two *dynamic* and *static* geo-based negative sampling approaches that take the sequential order of the users' checkins into account can generate effective top-K personalised venue recommendations. So far, we have investigated how the sequential order of the users' checkins play an important role in improving the quality of venue recommendations and *Recurrent Neural Networks* can effectively capture the users' *dynamic* preferences from their sequence of checkins. Although our proposed DRCF framework can effectively capture both the users' *dynamic* and *static* preferences from the users' sequence of checkins, it does not take the contextual information associated with the users' checkins into account. However, as mentioned in Section 1.1, to generate high quality Context-Aware Venue Recommendation (CAVR), users' contexts such as time of the day and the user's current location need to be considered. In the next chapter, we propose a recurrent neural network architecture that take the users' context into account to effectively generate high quality of CAVR.

Chapter 6

Contextual Attention Recurrent Architecture

6.1 Introduction

In the previous chapter, we described our proposed Deep Recurrent Collaborative Filtering framework (DRCF), which exploits the traditional RNN models to effectively capture the users' short-term (*dynamic*) preferences from the users' sequence of checkins. Our experimental results in the previous chapter demonstrate the effectiveness of the DRCF framework in generating effective top-K venue recommendations in comparison with various state-of-the-art MF-based approaches. On the other hand, as mentioned in Section 3.6.2, we argue that the traditional RNN models such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Chung et al., 2014) are limited as they can only take the sequential order of checkins into account and cannot incorporate the contextual information associated with the checkins (e.g. timestamp of a user's checkin and the geographical location of the checkin). Indeed, such contexts can influence the users' *dynamic* preferences, which have been shown to play an important role in producing effective Context-Aware Venue Recommendation (CAVR) (Liu et al., 2016c,b; Zhu et al., 2017). For example, a user is likely to visit a bar after he/she visited a restaurant. However, if the user visited a restaurant last night, recommending a bar to visit in the next day at daytime might not be appropriate. We argue that the traditional RNN-based factorisation approaches that do not take the contextual information associated with the users' checkins into account are likely to generate less effective venue recommendations than models that take these contextual information into account (**Limitation R3**, identified in Section 3.6.3). Table 6.1 provides the summary of these existing RNN-based factorisation approaches and their corresponding lim-

Table 6.1: Summary of the existing RNN-based factorisation approaches and the GRU architectures and their corresponding limitations

	RNN-MF	TimeGRU	CGRU	LatentCross	CARA
Sequential-based	✓	✓	✓	✓	✓
GRU Architecture	×	✓	✓	✓	✓
Ordinary Context	×	×	✓	✓	✓
Transition Context	×	only time	✓	✓	✓
Special Gates	×	✓	×	×	✓
Limitations	R3	G1-G4	G3-G4	G3-G4	-

itations. **Limitation R3** of the existing RNN-based factorisation approaches (e.g. RNN-MF) can be summarised below:

Limitation R3: There is a disadvantage in the RNN-based factorisation approaches that model the user's dynamic preferences from the sequential order of checkins by leveraging only the sequence of previously visited venues and ignoring the context associated with the checkins.

As described in Section 3.6.2, to incorporate the contextual information associated with the sequence of users' checkins, various Gate Recurrent Unit (GRU) architectures have been proposed, namely TimeGRU (Zhu et al., 2017), CGRU (Smirnova and Vasile, 2017) and LatentCross (Beutel et al., 2018). However, as discussed in Section 3.6.3, there are four limitations associated to these existing GRU architectures (**Limitations G1-G4**), with respect to how they treat the contextual information associated with the sequence of users' checkins. Indeed, we argue that there are two types of contextual information, namely the *ordinary* and *transition* contexts. For example, the ordinary context such as the time of the day should influence the user's contextual *dynamic* preferences on a current visited venue, while the transition context such as the time interval between the last visited venue and the current time should influence the correlation between the current and previously visited venues. Unfortunately, the TimeGRU architecture can only incorporate the time intervals between two successive checkins, (i.e. the transition context)) but not the current context of the users, (i.e. the ordinary context) (**Limitation G1**). In addition, the TimeGRU architecture is not sufficiently flexible to incorporate multiple types of the *transition* context associated with the checkins such as the geographical distance between two successive checkins (**Limitation G2**). Both CGRU and LatentCross are the existing state-of-the-art GRU architectures that aim to address **Limitation G1** by incorporating both the *ordinary* and *transition* contexts associated with the sequences of checkins. However, both CGRU and LatentCross treat the *ordinary* and *transition* contexts similarly, which contradict to our assumption, in which we assume that these two types of contexts influence the users' *dynamic* preferences differently (**Limitation G3**). Furthermore, there is a loss of granularity from the quantisation mapping functions used by both CGRU and LatentCross to represent the *transition* context (**Limitation G4**) [1]. Table 6.1 also provides a summary of these existing GRU architectures and their corresponding limitations. **Limitations G2-G4** of the existing GRU architectures are summarised as follows:

Limitation G2: The time gating mechanism proposed by TimeGRU is not sufficiently flexible to incorporate multiple types of transition contexts associated with the sequence of checkins.

Limitation G3: The GRU architectures for which this limitation applies (CGRU and LatentCross) treat the ordinary and transition contexts similarly, while these two types of context influence the users' preferences differently and should be treated independently.

Limitation G4: There is a disadvantage in the GRU architectures (CGRU and LatentCross) that rely on the quantised mapping procedures to represent transition contexts.

In this chapter, we aim to address **Limitations G2-G4** of the existing GRU architectures by proposing a novel Contextual Attention Recurrent Architecture (CARA) for context-aware venue recommendation that can effectively incorporate different types of contextual information associated with the users' sequence of checkins. In particular, our proposed CARA architecture consists of two types of gating mechanisms: namely a Contextual Attention Gate (CAG) as well as Temporal and Spatial Gates (TSG). The CAG gate aims to effectively capture the users' contextual *dynamic* preferences by taking into account the *ordinary* context associated with the users' checkins, while the TSG gates aim to capture the correlation between the users' previous checkin and the current checkin from the *transition* context associated with two successive checkins. The remainder of this chapter is structured as follows:

- Section 6.2 provides an overview of our proposed Contextual Attention Recurrent Architecture (CARA) with a *pairwise* ranking function that generates effective top-K venue recommendations to the users based on their sequence of historical checkins and context.
- Section 6.3 provides details of the first gating mechanism of the CARA architecture, the Contextual Attention Gate (CAG), which aims to capture the users' contextual *dynamic* preferences based on their *ordinary* context.
- Section 6.4 provides details of the second gating mechanism of the CARA architecture, the Temporal and Spatial Gates (TSG), which aims capture the correlation between the

[1] Further details have been already discussed in Section 3.6.2.2

users' previous checkin and the current checkin based on the users' *transition* context such as the time interval and distance between two successive checkins.

- Section 6.5 presents our experimental methodology in terms of datasets and measures as well as describes algorithm parameters.
- In Section 6.6, we empirically evaluate the effectiveness of our proposed CARA architecture in comparison with the state-of-the-art GRU architectures and RNN-based factorisation approaches, discussed in Chapter 3.
- Section 6.9 provides a summary of this chapter.

6.2 Overview of the Contextual Attention Recurrent Architecture (CARA)

In this section, we describe an overview of our proposed Contextual Attention Recurrent Architecture (CARA), an extension of the traditional GRU architecture that can incorporate both the *ordinary* and *transition* contexts associated with the users' sequence of checkins to effectively capture the users' short-term *dynamic* preferences. In particular, CARA aims to generate the ranked-list of venues that a user u might prefer to visit at time t based on his/her sequences of checkins $s_{u,t}$. Figure 6.1 illustrates the overview of the CARA architecture. The architecture consists of four layers: namely input, embedding, recurrent and output layers. The output of one layer serves as the input of the layer above. Starting at the bottom of the figure, at time step τ, the input layer consists of binary sparse vectors that use one-hot encodings to represent user u, venue i, and time t, respectively, as well as continuous values of time intervals and distances between two successive checkins. We compute the time interval and geographical distance between the given venue i and venue j previously visited at time step $\tau - 1$, as $\Delta t^\tau = t^\tau - t^{\tau-1}$ and $\Delta g_\tau = dist(lat_i, lng_i, lat_j, lng_j)$, respectively, where $lat_i, lng_i, lat_j, lng_j$ are the geographical location of venue i and j in terms of latitude and longitude, and $dist()$ is the Haversine distance function. The sparse vectors of the user u_u, venue v_i^τ and time t^τ are fed into the embedding layer. In the embedding layers, there are three latent factors of users $P \in \mathbb{R}^{m,d}$, venues $Q \in \mathbb{R}^{n,d}$ and times $M \in \mathbb{R}^{o,d}$, where m, n and o are the number of users, venues and time slots and d is the number of dimensions of the latent factors. The outputs of the embedding layer can be seen as the latent factors of user $\phi u_u = P^T u_u$, venue $\phi v_i^\tau = Q^T v_i^\tau$ and time $\phi t^\tau = M^T t^\tau$ in the context of a factorised model, which are denoted as the orange rectangles in Figure 6.1. $\theta_e = \{P, Q, M\}$ denotes the set of parameters of the embedding layer. Note that we only consider the time of checkins

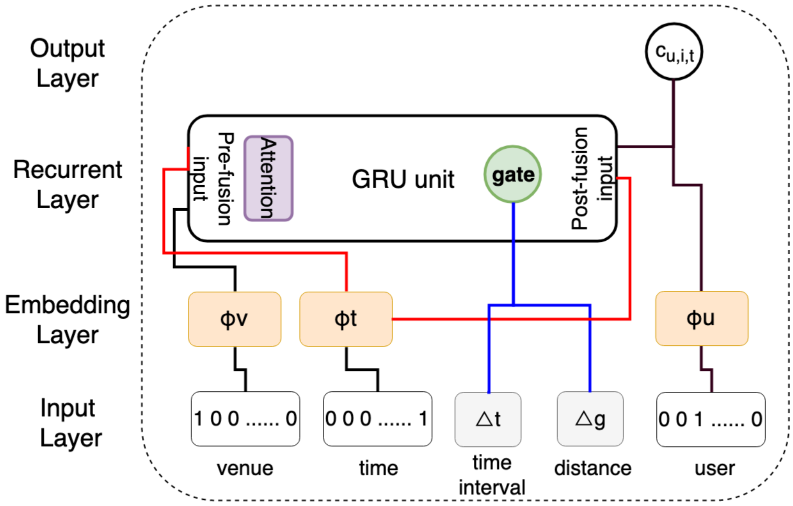


Figure 6.1: An overview of Contextual Attention Recurrent Architecture (CARA) unit at time step τ.

as the ordinary context but our proposed architecture is flexible to support multiple types of ordinary contexts (e.g. current weather of the day).

Next, the latent factors (embeddings) of venue ϕv_i^τ and time ϕt^τ as well as the time interval Δg_τ and the geographical distance Δt^τ are fed into the recurrent layer. The output of the recurrent layer is the hidden state of the recurrent unit at time step τ, h_τ, which is defined as follows:

$$h_\tau = f(\phi v_i^\tau, \phi t^\tau, \Delta t_\tau, \Delta g_\tau; \theta_r) \tag{6.1}$$

where $\theta_r = \{W, R, U, b\}$ denotes the set of parameters of the recurrent layer. We further describe the details of the recurrent units in the recurrent layer that generates the hidden state h_τ in Section 6.3 and Section 6.4. Finally, in the output layer, we estimate the preference of user u for venue i at timestamp t as follows:

$$\hat{c}_{u,i,t} = \phi u_i^T h_\tau \tag{6.2}$$

where $h_\tau \in \mathcal{R}^d$ is the hidden state of the recurrent layer. We apply the pairwise Bayesian Personalised Ranking (BPR) (Rendle et al., 2009) to learn the parameters $\Theta = \{\theta_e, \theta_r\}$, as follows:

$$\mathcal{J}(\Theta) = \sum_{i \in \mathcal{U}} \sum_{s_{u,t} \in S_u} \sum_{(u,i,t) \in s_{u,t}} \sum_{j \in \mathcal{V} - s_{u,t}} \log(\sigma(\hat{c}_{u,i,t} - \hat{c}_{u,j,t})) \tag{6.3}$$

where j is a venue the user has not visited before up to time t. Finally, the gradients of θ_r and θ_e can be estimated by back-propagation through time algorithm proposed by Rumelhart et al. (1988). In the next section, we describe the first gating mechanism of the CARA archi-

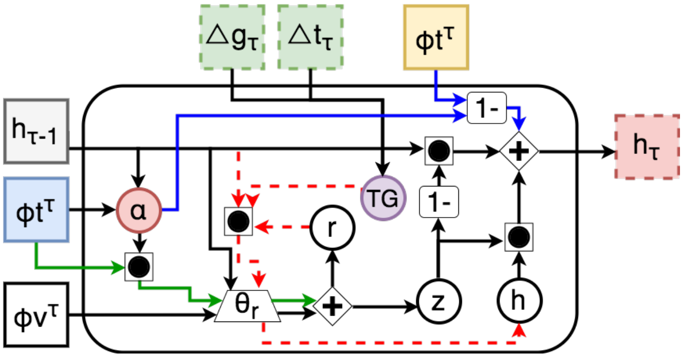


Figure 6.2: The gating mechanisms of our proposed Contextual Attention Recurrent Architecture (CARA). The Rectangle symbols indicate inputs of the unit, a red-dashed rectangle symbol indicates the output of the unit and the circle symbols are the unit' gates. This figure is obtained from Manotumruksa et al. (2018)

tecture, the Contextual Attention Gate, which aims to capture the users' contextual *dynamic* preferences based on the *ordinary* context associated with the sequence of users' checkins (i.e. the time of the checkins).

6.3 Contextual Attention Gate (CAG)

As mentioned in Section 6.1, we argue that the *ordinary* and *transition* contexts associated with the users' checkins can influence the users' contextual *dynamic* preferences differently. Indeed, the *ordinary* context (e.g. time t) influences the user's preference on a current visited venue, while the transition context (e.g. the time interval ΔT) influences the correlation between the current and previously visited venues. In this section, we describe how we extend the gating mechanism of the traditional Gated Recurrent Unit (GRU), described in Section 2.2.3.3, to effectively incorporate the ordinary context associated with the sequences of users' checkins. In particular, we further describe how to calculate the hidden state h_τ in Equation (6.1).

Figure 6.2 illustrates the gating mechanisms of our proposed CARA architecture. First, inspired by Donkers et al. (2017), we propose the Contextual Attention Gate (CAG), $\alpha \in \mathbb{R}^d$, denoted as the red circle in Figure 6.2, which controls the influences of the latent factor of time ϕt^τ at each time step τ as follows:

$$\alpha_\tau = \sigma(W_{\alpha,h} h_{\tau-1} + W_{\alpha,t} \phi t^\tau + b_\alpha) \tag{6.4}$$

where $W_{\alpha,h}$ and $W_{\alpha,t}$ are the weight matrix of the attention gate for the latent factors of time ϕt^τ and the hidden state $h_{\tau-1}$ of the previous unit, respectively, and b_α is a bias parameter.

The attention gate α_τ aims to capture the correlation between the latent factor ϕt^τ at current time step τ and the hidden state $h_{\tau-1}$ of the previous unit. Then, we modify the gating mechanism of the traditional GRU (Equations (2.29)-(2.32)) by including the attention gate α_τ as follows:

$$[z_\tau, r_\tau] = \sigma(W\phi v_i^\tau + Rh_{\tau-1} + W(\alpha_\tau \odot \phi t^\tau) + b) \tag{6.5}$$

$$\widetilde{h}_\tau = \tanh(W\phi v_i^\tau + R(r_\tau \odot h_{\tau-1}) + W(\alpha_\tau \odot \phi t^\tau) + b) \tag{6.6}$$

$$h_\tau = (1 + (1 - \alpha_\tau) \odot \phi t^\tau) \odot [(1 - z_\tau)h_{\tau-1} + z_\tau \widetilde{h}_\tau] \tag{6.7}$$

Note that, unlike CGRU (Beutel et al., 2018) and LatentCross (Smirnova and Vasile, 2017), described in Section 3.6.2.2, which combine the latent factors of the venue ϕv_i^τ and time ϕt^τ using the concatenation operation, i.e. $x^\tau = [\phi v_i^\tau; \phi t^\tau]$ in Equation (3.28), we argue that these two latent factors should be treated independently. Ideally, at time step τ, the ordinary context associated with the checkins (i.e. the latent factors of time ϕt^τ) represents the user's contextual preferences about the venue, while the latent factors of the venue ϕv_i^τ represent the characteristics of the venues. Indeed, we can include the ordinary context into the GRU units in two ways: namely at the beginning and the end of the GRU unit. Cui et al. (2010) described the inclusion of context features before the GRU unit as *pre-fusion* (blue box in Figure 6.2), and the inclusion of context features after the GRU unit as *post-fusion* (yellow box in Figure 6.2). In particular, by including the latent factors of time ϕt^τ through pre-fusion (Equations (6.5) & (6.6)), ϕt^τ will affect the update of the hidden state of the current GRU unit though the update and reset gates z_τ, r_τ as well as the candidate hidden state $\widetilde{h}_\tau$. However, by including the latent factors of time ϕt^τ through post-fusion (Equation (6.7)), ϕt^τ has more effect on the hidden state h_τ, the output of the current GRU unit, and hence affects the next hidden state of next step $h_{\tau+1}$. Our proposed attention gate α_τ controls the influence of the latent factor of time ϕt^τ on both pre- and post- fusion.

6.4 Time- and Spatial-based Gates (TSG)

In the previous section, we have explained how to extend the gating mechanism of the traditional GRU unit to incorporate the ordinary context associated with the users' observed checkins. As mentioned in Section 6.1, to effectively model the users' sequential order of checkins, we also need to take the transition context associated with the users' observed checkins into account. In this section, we describe how to further extend the gating mechanism of the traditional GRU unit to incorporate the transition context such as the time intervals (i.e. Δt^τ) and the geographical distances (i.e. Δg_τ) between successive checkins. In Figure 6.2, the purple circle illustrates our proposed Time- and Spatial-based Gates (TSG), while the green-dashed boxes illustrate the inputs of the TSG gates. Indeed, our proposed TSG gates together with the CAG gate, previously described in Section 6.3, aim to address **Limitations G2-G4** of the existing gating mechanisms (TimeGRU, CGRU and LatentCross). In particular, inspired by the time gate proposed by Zhu et al. (2017) (TimeGRU), described in Section 3.6.2.1, we propose to extend their time gate to incorporate both the time interval and geographical distance between two checkins as follows:

$$T_\tau = \sigma_t(W_{tx}\phi v_i^\tau + \sigma(\Delta t_\tau W_t) + b_t) \tag{6.8}$$

$$G_\tau = \sigma_t(W_{gx}\phi v_i^\tau + \sigma(\Delta g_\tau W_g + b_g)) \tag{6.9}$$

where Δt^τ and Δg_τ are the time interval and the distance between checkins c_τ and $c_{\tau-1}$, at time step τ, respectively. W_t and W_g are the weight matrix for each transition context (i.e. the time interval and geographical distance, respectively). Similarly, b_t and b_g are the bias parameters. Note that unlike the CGRU and LatentCross architectures, our proposed TSG gates support using continuous values for a transition context, hence they do not rely on the quantised mapping procedure to represent a transition context, while CGRU and LatentCross do (**Limitation G4**). In fact, by considering the continuous values of the transition context, the TSG gates are more effective than CGRU and LatentCross in capturing the correlation between the users' previous checkin and the current checkin based on the transition context between these two successive checkins.

Next, we propose to combine these two gates, T_τ and G_τ, using the element-wise product $TG_\tau = T_\tau \odot G_\tau$ and modify Equation (6.6) as follows:

$$\widetilde{h}_\tau = \tanh(W\phi v_j^\tau + R(r_\tau \odot TG_\tau \odot h_{\tau-1}) + W(\alpha_\tau \odot \phi t^\tau) + b) \tag{6.10}$$

The TG_τ gate and the reset gate r_τ together control the influence of the hidden state of the previous step $h_{\tau-1}$. Unlike the TimeGRU architecture (**Limitation G2**), the TG_τ gates can effectively take both the time interval and the geographical distance of two successive checkins into account. Hence, even if the time interval between two checkins is long, the influence of the hidden state from the previous time step $h_{\tau-1}$ may not be decreased if the distance between the two checkins is short. Based on the assumption we mentioned in Section 3.6.2, for example, a user who visited a museum yesterday is likely to visit another museum nearby the visited museum, although the time interval from the previous checkin is long. Later in Section 6.6, we demonstrate that our proposed TSG gates are more effective than the time gate proposed by Zhu et al. (2017) (TimeGRU) in capturing the correlation between the two successive checkins based on the transition context. Finally, to address **Limitation G3** of the CGRU and Latent Cross architectures, together by combining the CAG and TSG gates, our proposed CARA architecture can effectively capture the users' contextual *dynamic* preferences by treating the ordinary and transition contexts associated with the sequence of users' checkins independently. In the next section, we describe our experimental methodology used to evaluate the effectiveness of our proposed CARA architecture.

6.5 Experimental Methodology

In the previous section, we described our proposed Contextual Attention Recurrent Architecture (CARA), which consists of two gating mechanisms: namely the Contextual Attention Gate (CAG) and the Time- and Spatial-based Gates (TSG). Our proposed CARA architecture aims to address **Limitations G2-G4** of the existing Gated Recurrent Unit (GRU) architectures (see Section 3.6.3). In this section, we evaluate the effectiveness of our proposed CARA architecture in comparison with state-of-the-art GRU architectures and factorisation approaches. In particular, to address **Limitations G2-G4**, we aim to answer the following research questions:

RQ6.1 *Can we enhance the effectiveness of the traditional recurrent architecture (GRU) by leveraging the ordinary and transition contexts associated with the sequence of checkins?*

RQ6.2 *Is it important to model the ordinary and transition contexts separately?*

RQ6.3 *Does the use of the absolute continuous values of the transition context preserve the influence of successive checkins?*

Furthermore, as discussed in Section 3.6.3, no previous work has proposed a gating mechanism that can incorporate multiple types of transition contexts such as the time intervals and geographical distances between two successive checkins. Hence, we initiate our final research question as follows:

RQ6.4 *Can our proposed Time- and Spatial-based Gates (TSG) that leverage multiple types of transition contexts enhance the effectiveness of traditional recurrent units in capturing the user's contextual dynamic preferences?*

Table 6.2 summarises the research questions we aim to address in this chapter and their corresponding success decisions. In particular, we aim to answer research question RQ6.1 by comparing the performance of CARA with the existing factorisation approaches as well as the existing GRU architectures. Next, by answering research question RQ6.2,

Table 6.2: Summary of each research question and its corresponding success decision and the limitations of the existing approaches.

Research Question	Limitation	Success Decision
RQ6.1	-	Our proposed CARA architecture that takes both the *ordinary* and *transition* contexts into account is more effective than the existing GRU architectures and factorisation approaches.
RQ6.2	G3	Our proposed CARA architecture that treats the *ordinary* and *transition* contexts separately via its gating mechanisms (CAG and TSG gates) is more effective in generating high quality top-K CAVR than the existing GRU architectures (CGRU and LatentCross) that treat both *ordinary* and *transition* contexts equally.
RQ6.3	G4	Our proposed TSG gates of the CARA architecture that rely on the absolute continuous values of the *transition* context are more effective in capturing the users' contextual *dynamic* preferences than CGRU and LatentCross, which rely on the quantised mapping procedures to represent the transition context.
RQ6.4	G2	Our proposed TSG gates of the CARA architecture are more effective than the time gate of TimeGRU when multiple types of *transition* contexts associated with the users' sequence of checkins can be observed.

we can demonstrate that our proposed CAG and TSG gates of the CARA architecture can address **Limitation G3** of the state-of-the-art GRU architectures (CGRU and LatentCross). Hence, we can also show that the *ordinary* and *transition* contexts are different and should be treated separately. Furthermore, by answering research question RQ6.3, we can demonstrate that our proposed TSG gates of the CARA architecture, which rely on the absolute continuous values of the *transition* context, can address **Limitation G3** of the CGRU and LatentCross architectures. Finally, by answering research question RQ6.4, we can show that our proposed TSG gates can address **Limitation G2** of TimeGRU by incorporating multiple types of *transition* contexts.

In the remainder of this section, we describe the experimental setup in terms of datasets and measures (Section 6.5.1), baselines (Section 6.5.2) and the algorithm parameters (Section 6.5.3). The experimental results and analysis follow in Section 6.6.

Table 6.3: Statistics of the three used datasets.

	Brightkite	Foursquare	Yelp
Number of normal users	14,374	10,766	38,945
Number of venues	5,050	10,695	34,245
Number of ratings or checkins	681,024	1,336,278	981,379
Number of cold-start users	5,578	154	6903
% density of User-Venue matrix	0.93	1.16	0.07

6.5.1 Datasets & Measures

Similar to the experimental setup described in Chapter 5, we conduct experiments using three publicly available large-scale LBSN checkin and rating datasets. In particular, to show the generalisation of our proposed CARA architecture across multiple LBSN platforms and sources of feedback evidence, we use two checkin datasets from Brightkite[2] and Foursquare[3], and a rating dataset from Yelp[4]. We follow the common practice from previous works (Rendle et al., 2009; He et al., 2017) to remove venues with less than 10 checkins. Table 6.3 summarises the statistics of the filtered datasets. To evaluate the effectiveness of our proposed CARA architecture and following previous works (He et al., 2017; Rendle et al., 2009), we adopt a *leave-one-out* evaluation methodology: for each user, we select their most recent checkin as a ground truth and randomly select 100 venues that they have not visited before as the testing set, where the remaining checkins are used as the training set. The context-aware venue recommendation task is thus to rank those 101 venues for each user given their preferred context (i.e. time), aiming to rank highest the recent, ground truth checkin. We conduct two separate experiments, namely: *Normal Users* (those with ≥ 10 checkins) and *Cold-start Users* (< 10 checkins) to evaluate the effectiveness of our proposed CARA architecture in the general and cold-start settings. Recommendation effectiveness is measured in terms of Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) on the ranked lists of venues – as applied in the literature (He et al., 2017; Yu et al., 2016). In particular, HR considers the ranking nature of the task, by taking into account the rank(s) of the venues that each user has previously visited/rated in the produced ranking, while NDCG goes further by considering the checkin frequency/rating value of the user as the graded relevance label. Finally, significance tests are conducted using a paired t-test.

[2] https://snap.stanford.edu/data/

[3] https://archive.org/details/201309_foursquare_dataset_umn

[4] https://www.yelp.com/dataset_challenge

6.5.2 Baselines

We compare our proposed Contextual Attention Recurrent Architecture (CARA) with various baselines, which can be categorised as the trivial sanity-check baselines, the state-of-the-art recurrent neural network architectures and the factorisation approaches. Note that some approaches and frameworks may not be originally proposed for Context-Aware Venue Recommendation (CAVR) but are sufficiently flexible to be applied to this task without any disadvantages. We implement all baselines and our proposed CARA architecture using Keras[5], a deep learning framework built on top of Theano[6]. Our implementations are released as open source[7]. For a fair comparison, the choice of recurrent models is fixed to the Gated Recurrent Units (GRU) proposed by (Chung et al., 2014), described in Section 2.2.3.3. In addition, compared to LSTM, GRU has less parameters yet is as effective as LSTM for recommendations(Smirnova and Vasile, 2017; Tan et al., 2016; Tang et al., 2017). The summary of the baselines are described below.

6.5.2.1 Trivial Sanity-Check Baselines

MostPop is a baseline that ranks venues in descending order of the venues' popularities, calculated across all users.

MostVisit is a baseline that ranks venues for a given user in descending order of the venues' popularity for that user.

RecentVisit is a baseline that takes the user's sequential order of checkins into account and recommends the most recently visited venue to the user.

MF (Koren et al., 2009) is the traditional matrix factorisation approach that aims to accurately predict the users' checkin on the unvisited venues.

6.5.2.2 Recurrent Neural Network Architectures

RNN-MF (Zhang et al., 2014b) is a traditional recurrent architecture that only takes the sequence of users' checkins into account and ignores the contextual information associated with the sequence of checkins (see Section 3.6.1).

STGRU (Liu et al., 2016c) is a Spatial and Temporal recurrent architecture that incorporates multiple types of transition contexts associated with the sequence of users' checkins (i.e. the time intervals and the geographical distances between checkins).

[5] https://github.com/fchollet/keras

[6] http://deeplearning.net/software/theano

[7] https://github.com/feay1234/CARA

CAGRU (Liu et al., 2016b) is an extension of the STGRU architecture that can incorporate both the ordinary and transition contexts associated with the sequence of users' checkins.

TimeGRU (Zhu et al., 2017) is an extension of the GRU architecture that includes the time gate to incorporate a time interval between successive checkins (see Section 3.6.2.1).

CGRU (Smirnova and Vasile, 2017) is an extension of the GRU architecture that can incorporate multiple types of contexts. (see Section 3.6.2.2).

LatentCross (Beutel et al., 2018) is an extension of CGRU that supports pre- and post-fusion inputs (see Section 3.6.2.2).

6.5.2.3 Factorisation Approaches

MF (Koren et al., 2009) is the traditional matrix factorisation approach that aims to accurately predict the users' checkin on the unvisited venues based on their historical checkins (see Section 2.1.1.1.2).

BPR (Rendle et al., 2009) is the classical pairwise ranking approach, coupled with matrix factorisation for user-venue checkin prediction (see Section 2.1.1.1.2).

GeoBPR (Yuan et al., 2016) is the extension of BPR that incorporates the geographical location of venues to effectively sample negative venues that are far away from the user's previous visits (see Section 3.5.1).

STELLAR (Zhao et al., 2016) is a Spatial-TEmporaL LAtent Ranking framework for CAVR that aims to recommend the list of venues based on the user's preferred time and last successive visits. Note that this is the only context-aware venue recommendation framework that does not rely on the RNN-based models to model the users' sequential order of checkins.

NeuMF (He et al., 2017) is the Neural Matrix Factorisation framework[8] that exploits Multi-Level Perceptron (MLP) models to capture the complex structure of user-venue interactions (see Section 3.4).

DRCF is our proposed Deep Recurrent Collaborative Filtering framework, previously described in Chapter 5, which extends the NeuMF framework to exploit the traditional RNN models to capture the sequential order of users' checkins. Recall that the DRCF framework consists of two components, with each component having its own recurrent layer. Hence, to permit a fair comparison, we only compare our proposed CARA architecture with its best-performing component, the GRMF model (see Section 5.4), which uses an element-wise product to combine the latent factors and the hidden units of the RNN model.

[8] `https://github.com/hexiangnan/neural_collaborative_filtering`

6.5.3 Recommendation Parameters Setup

In this section, we describe how we set the parameters of our proposed CARA architecture and baselines to permit a fair comparison. Following (He et al., 2017), we set the dimensions of the latent factors d and hidden layers h_τ of the CARA architecture and all of the factorisation approaches to be identical: $d = 10$ across the three datasets. Following He et al. (2017), we randomly initialise all embeddings and recurrent layers' parameters, θ_e, θ_r, with a Gaussian distribution (with a mean of 0 and a standard deviation of 0.01) and apply the mini-batch Adam optimiser (Kingma and Ba, 2014) to optimise those parameters, which yields a faster convergence than SGD and automatically adjusts the learning rate for each iteration. We initially set the learning rate to 0.001[9] and set the batch size to 256. As the impact of the recurrent parameters such as the size of the hidden state, have been explored in previous works (He et al., 2017, 2016b; Tan et al., 2016), we omit varying the size of the hidden layers and the number of dimensions of the latent factors in this work. Indeed, larger sizes of hidden layers and dimensions may cause over-fitting and degrade the generalisation of the models (He et al., 2016b, 2017; Tan et al., 2016).

6.6 Experimental Results

In this section, to answer research questions RQ6.1-RQ6.4 identified in Section 6.5, we evaluate the effectiveness of our proposed CARA architecture in comparison with various baselines described in Section 6.5.2. In particular, to address research questions RQ6.1-RQ6.3, we compare the CARA architecture with the state-of-the-art GRU architectures as well as factorisation approaches under the *Normal Users* and *Cold-Start Users* settings in Section 6.7. Next, to address research question RQ6.4, we further investigate the usefulness of the gating mechanisms of the CARA architecture in enhancing the quality of the top-K context-aware venue recommendations in several settings in Section 6.8

6.7 The Effectiveness of CARA Architecture

Table 6.4 reports the effectiveness of our proposed CARA architecture and various state-of-the-art GRU recommendation architectures, in terms of HR@10 and NDCG@10 on the three used datasets. This table consists of two groups of rows, which report the effectiveness of various approaches under the *Normal Users* and *Cold-Start Users* settings, respectively.

[9] The default learning rate setting of the Adam optimiser in Keras.

Table 6.4: Performances in terms of HR@10 and NDCG@10 of various approaches. The best performing result is highlighted in bold; $-$ and $*$ denote a significant difference compared to the best performing result, according to the paired t-test for $p < 0.05$ and $p < 0.01$, respectively.

	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
Normal Users						
MostPop	0.1462*	0.1010*	0.2009*	0.1167*	0.0739*	0.0334*
MostVisit	0.4032*	0.3473*	0.4733*	0.4290*	0.1083*	0.0528*
RecentVisit	0.4809*	0.4370*	0.4584*	0.4037*	0.1096*	0.0542*
RNN-MF	0.6657*	0.4407*	0.8302*	0.5762*	0.4164*	0.2146*
TimeGRU	0.7005*	0.4816*	0.8570*	0.6167*	0.4342*	0.2240*
STGRU	0.6888*	0.5493*	0.8496*	0.6865*	0.4254*	0.2365*
CAGRU	0.7180*	0.5545*	0.8498*	0.6474*	0.3799*	0.1989*
CGRU	0.6969*	0.5659*	0.8592*	0.6985*	0.5194*	0.3005*
LatentCross	0.7063*	0.5727*	0.8616*	0.6964*	0.5210*	0.2991*
CARA	**0.7385**	**0.6040**	**0.8851**	**0.7154**	**0.5587**	**0.3272**
Cold-Start Users						
MostPop	0.1155*	0.0778*	0.0584*	0.0286*	0.0714*	0.0316*
MostVisit	0.4285*	0.3789*	0.3506*	0.3175*	0.1044*	0.0489*
RecentVisit	0.4995*	0.4585*	0.3831*	0.3446*	0.1052*	0.0497*
RNN-MF	0.6959*	0.4550*	0.8247	0.5260*	0.2420*	0.4540*
TimeGRU	0.7314*	0.5071*	0.8182	0.5788*	0.2398*	0.4592*
STGRU	0.7081*	0.5686*	0.7273*	0.5722*	0.2543*	0.4404*
CAGRU	0.7628	0.6035*	0.8377	0.6353	0.2205*	0.4055*
CGRU	0.7054*	0.5788*	0.7662*	0.5996-	0.3325*	0.5524*
LatentCross	0.7108*	0.5811*	0.8052*	**0.6600**	0.3223*	0.5398*
CARA	**0.7648**	**0.6220**	**0.8636**	0.6505	**0.3493**	**0.5748**

Table 6.5 reports the mean percentage improvements of our proposed CARA architecture over various GRU baseline architectures across the three used datasets. These percentage improvements are reflected from the experimental results reported in Table 6.4. Similar to Table 6.4, Table 6.5 consists of two groups of rows, which report the percentage improvements of CARA over various approaches under the *Normal Users* and *Cold-Start Users* experiments, respectively. Firstly, on inspection of our reimplementations of the state-of-the-art GRU baselines, we note that the relative venue recommendation qualities of the baselines on the three used datasets in terms of both HR and NDCG are consistent with the results reported for the various baselines in the corresponding literature (Zhang et al., 2014b; Beutel et al., 2018; Smirnova and Vasile, 2017; Liu et al., 2016c,b; Zhu et al., 2017). For instance, the extensions of the GRU architectures that incorporate the contextual information (LatentCross, CGRU, CAGRU, STGRU and TimeGRU) outperform RNN-MF across the three used datasets.

Table 6.5: Mean percentage improvements of CARA over various baselines across the three used datasets in terms of HR@10 and NDCG@10, which are obtained from Table 6.4.

Normal Users		
Model	HR	NDCG
RNN	17.24%	37.89%
TimeGRU	12.46%	29.16%
STGRU	14.24%	17.51%
CAGRU	18.02%	27.98%
CGRU	5.52%	6.01%
Latent Cross	4.84%	5.86%
Cold-Start Users		
Model	HR	NDCG
RNN	19.65%	28.99%
TimeGRU	18.59%	20.07%
STGRU	21.37%	9.78%
CAGRU	20.59%	15.74%
CGRU	8.73%	6.67%
Latent Cross	7.74%	4.03%

Within the first group of rows in Table 6.4 and Table 6.5, comparing CARA with various GRU architectures on the Normal Users experiment, we observe that CARA consistently and significantly outperforms all the GRU baselines, for HR and NDCG, across all datasets. In particular, on the average across the three used datasets, CARA improves HR and NDCG by 5-18% and 6-27%, respectively, over the recently proposed GRU architectures, CAGRU, CGRU and LatentCross (see Table 6.5). These results suggest that our proposed CARA architecture with Contextual Attention Gate (CAG) and Time-and Spatial-based Gates (TSG) is more effective than the state-of-the-art GRU architectures in modelling the sequences of users' checkins. Therefore, in response to research question RQ6.1, the experimental results reported in Tables 6.4 and 6.5 demonstrate that the effectiveness of a recurrent architecture in capturing the users' contextual *dynamic* preferences can be further improved by taking the ordinary and transition contexts associated with the sequence of checkins into account.

Next, to answer research questions RQ6.2 and RQ6.3, we compare CARA with the GRU baseline architectures that consider both the ordinary and transition contexts (CAGRU, CGRU and LatentCross). Note that these GRU baselines treat the ordinary and transition contexts similarly and rely on the quantised mapping procedures to represent the contexts. However, as mentioned in Section 6.1, we argue that different types of contexts might influence the user's dynamic preferences differently. In addition, using the mapping procedure to convert the continuous values of the transition context can lead to a loss in granularity. From the experimental results in Table 6.4, we observe that our proposed CARA architecture, which leverages the absolute continuous values of the transition context (i.e. the time interval Δt_τ and the geographical distance Δg_τ – see Section 6.4) is more effective than the CAGRU, CGRU and LatentCross baselines in capturing the transition effects between successive checkins. In particular, our proposed Contextual Attention Gate (CAG) enables the CARA architecture to treat the ordinary and transition contexts separately, while these GRU baselines do not.

Within the second group of rows in Tables 6.4 and 6.5, we further analyse the effectiveness of our proposed CARA architecture by comparing with the GRU baselines in the *Cold-Start Users* setting. Similar to the results observed from the *Normal Users* experiment, CARA consistently and significantly outperforms all GRU baselines across the three used datasets in terms of HR and NDCG, except for NDCG on the Foursquare dataset, where LatentCross is statistically indistinguishable from CARA (difference in HR $< 1.5\%$ $(0.6505 \rightarrow 0.6600)$). Next, we note that unlike the Brightkite and Foursquare checkin datasets, the Yelp dataset consists of only user-venue ratings, and hence the sequential properties of visits to venues cannot be observed. Consequently, in both normal and cold-start user settings, the performances of several GRU baselines (TimeGRU, STGRU and CAGRU) that consider the contextual information of the ratings are as effective as the RNN baseline that only considers the sequence of the user's ratings. In contrast, our proposed CARA architecture, which controls the influence of previous ratings based on both the time interval and the geographical distance, is still the most effective across the different types of datasets.

Next, to further response to research question RQ6.1, we investigate the performance of our proposed CARA architecture in comparison with the state-of-the-art factorisation approaches, described in Section 6.5.2.3. Table 6.6 reports the effectiveness of CARA and various factorisation approaches, in terms of HR@10 and NDCG@10 on the three used datasets. Similar to Table 6.4, Table 6.6 contains two groups of rows, which report the effectiveness of various approaches under the *Normal Users* and *Cold-Start Users* settings, respectively. Table 6.7 reports the mean percentage improvements of our proposed CARA architecture over various factorisation baselines across the three used datasets. These percentage improvements are summarised from the experimental results reported in Table 6.6. On inspection of our reimplementations of the factorisation baselines, we note that the relative venue recommendation qualities of the baselines on the three used datasets in terms of both HR and NDCG are consistent with the results reported for the various baselines in the corresponding literature (Zhao et al., 2016; He et al., 2017; Yuan et al., 2016). In particular, we observe the relative improvements of GeoBPR, STELLAR, NeuMF and DRCF compared to MF and BPR across the three datasets. Note that, indeed, while previous works (Zhao et al., 2016; He et al., 2017; Yuan et al., 2016) used different datasets, our reimplementations of their

Table 6.6: As per Table 6.4; comparison between our proposed CARA architecture and various factorisation baselines.

	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
Normal Users						
MostPop	0.1462*	0.1010*	0.2009*	0.1167*	0.0739*	0.0334*
MostVisit	0.4032*	0.3473*	0.4733*	0.4290*	0.1083*	0.0528*
RecentVisit	0.4809*	0.4370*	0.4584*	0.4037*	0.1096*	0.0542*
MF	0.6206*	0.3470*	0.6656*	0.3818*	0.3539*	0.1734*
BPR	0.6890*	0.4333*	0.7550*	0.4834*	0.4992*	0.2691*
GeoBPR	0.7339	0.4672*	0.8216*	0.5395*	0.5570	0.3020*
STELLAR	0.7267*	0.5635*	0.8751*	0.6984*	0.5356*	0.2969*
NeuMF	0.7073*	0.5358*	0.8361*	0.5842*	0.4927*	0.2734*
DRCF	0.7363	0.5670*	0.8805	0.6814*	0.5209*	0.2890*
CARA	**0.7385**	**0.6040**	**0.8851**	**0.7154**	**0.5587**	**0.3272**
Cold-Start Users						
MostPop	0.1155*	0.0778*	0.0584*	0.0286*	0.0714*	0.0316*
MostVisit	0.4285*	0.3789*	0.3506*	0.3175*	0.1044*	0.0489*
RecentVisit	0.4995*	0.4585*	0.3831*	0.3446*	0.1052*	0.0497*
MF	0.6768*	0.3913*	0.6623*	0.3650*	0.3748*	0.1868*
BPR	0.7519	0.4907*	0.7792-	0.4961*	0.5273*	0.2946*
GeoBPR	**0.8093**	0.5262*	0.8312	0.5486*	**0.5802**	0.3202*
STELLAR	0.7406*	0.5580*	0.8052-	0.6007-	0.5537*	0.3147*
NeuMF	0.7160*	0.5894*	0.7922-	0.6227	0.5102*	0.2956*
DRCF	0.7409*	0.5618*	0.8442	**0.6542**	0.5399*	0.3083*
CARA	0.7648	**0.6220**	**0.8636**	0.6505	0.5748	**0.3493**

factorisation approaches obtain similar relative improvements.

From the first group of rows in Table 6.6, we observe that CARA consistently and significantly outperforms all the factorisation baselines across the three used datasets in terms of HR and NDCG. In particular, comparing with STELLAR, the state-of-the-art CAVR that considers both the contextual information and the sequences of users' checkins, CARA obtains 7.19% ($0.5635 \rightarrow 0.6040$) and 10.21% ($0.2969 \rightarrow 0.3272$) improvements in terms of NDCG on the Brightkite and Yelp datasets, respectively. The percentage improvements of CARA over STELLAR on the average across the three used datasets in terms of HR and NDCG are 2.36% and 6.61%, respectively (see Table 6.7). In addition, comparing with our proposed DRCF framework, which exploits RNN models to capture the users' *dynamic* preferences (see Chapter 5), our proposed CARA architecture significantly outperforms DRCF by 6.53% ($0.5670 \rightarrow 0.6040$), 5% ($0.6814 \rightarrow 0.7154$) and 13.22% ($0.2890 \rightarrow 0.3272$) in terms of NDCG on the Brightkite, Foursquare and Yelp datasets, respectively. The mean percentage improvements of CARA over DRCF across the three datasets in terms of HR and NDCG are 2.69% and 8.24%, respectively (see Table 6.7). However, there is no significant

Table 6.7: Mean percentage improvements of CARA over various baselines across the three used datasets in terms of HR@10 and NDCG@10, which are summarised from Table 6.6.

Normal Users		
Model	HR	NDCG
MF	36.61%	83.38%
BPR	12.11%	36.33%
GeoBPR	2.89%	23.41%
STELLAR	2.36%	6.61%
NeuMF	7.89%	18.29%
DRCF	2.69%	8.24%
Cold-Start Users		
Model	HR	NDCG
MF	32.25%	74.72%
BPR	7.19%	25.48%
GeoBPR	-0.84%	15.29%
STELLAR	4.78%	10.25%
NeuMF	9.50%	9.39%
DRCF	4%	7.82%

improvement between DRCF and CARA in terms of the HR metric for the Brightkite and Foursquare datasets. In addition, we highlight that GeoBPR uses an advanced geo-based negative sampling technique (see Section 3.5.1), while CARA uses the traditional negative sampling, similar to BPR. CARA is as effective as GeoBPR in terms of HR on Brightkite and Yelp (i.e. no significant differences are observed), while using a less advanced sampling technique. We also underline that CARA can be adapted to use GeoBPR's negative sampling as well as our proposed *dynamic* and *static* geo-based negative sampling approaches (see Section 5.3), which we further discuss in Chapter 7.

From the second group of rows in Table 6.6, we observe that CARA consistently and significantly outperforms all the factorisation baselines across the Brightkite and Yelp datasets in terms of NDCG. In particular, comparing with STELLAR, the percentage improvements of CARA over STELLAR on the average across the three used datasets in terms of HR and NDCG are 4.78% and 10.25%, respectively (see Table 6.7). In addition, the mean percentage improvements of CARA over DRCF across the three datasets in terms of HR and NDCG are 4% and 7.82%, respectively. These results demonstrate that CARA is more effective than the state-of-the-art STELLAR and our proposed DRCF framework in alleviating the cold-start problem. Although, GeoBPR can outperform CARA in terms of HR on the Brightkite and Yelp datasets, there are no significant improvements between GeoBPR and CARA. Overall, in response to research question RQ6.1, the experimental results in Table 6.6 and Table 6.7 show that our proposed CARA architecture, which leverages the sequences of users' checkins as well as the ordinary and transition contexts associated with the checkins, is effective for CAVR. Furthermore, the experimental results also show that our proposed CARA architecture is more effective than various existing factorisation baselines in alleviating the cold-start problem.

6.8 The Usefulness of Gating Mechanisms of the CARA Architecture

In the previous section, we evaluated the effectiveness of our proposed CARA architecture in comparison with various baselines in order to address research questions RQ6.1 - RQ6.3. In this section, to further address the research questions RQ6.3-RQ6.4, we investigate the effectiveness of our proposed CARA architecture and the existing GRU baselines under different settings. In particular, Figure 6.3 presents the performances of various GRU architectures on the three used datasets in terms of HR@10 and NDCG@10, by considering the users with particular time intervals Δt (hours) and geographical distances Δg (km) between their last checkin and the ground-truth checkin. Overall, the experimental results observed in Figure 6.3 demonstrate that CARA can consistently outperform all the baselines in different settings across the three used datasets in terms of HR and NDCG.

With respect to research question RQ6.4, we first compare the performance of CARA and TimeGRU, the existing GRU architecture that consists of the time gate, described in Section 3.6.2.1. From Figure 6.3, we observe that CARA consistently outperforms TimeGRU in terms of HR and NDCG across the three used datasets on every time intervals Δt (hours) and geographical distances Δg (km). These results suggest that our proposed CARA architecture, which consists of Time- and Spatial-based Gates (TSG), is more effective than TimeGRU, the GRU baseline that considers only the time intervals. Therefore, by considering both the time interval and the geographical distance between two successive checkins, CARA can generate better recommendations than TimeGRU. Next, to further address research question RQ6.3, we compare CARA with CGRU and LatentCross, the GRU baselines that rely on the quantised mapping procedures to represent the transition contexts (**Limitation G4**). The experimental results from Figure 6.3 on the three used datasets demonstrates that our proposed CARA architecture, which supports the absolute continuous values of the *transition* contexts by using the TSG gates, outperforms CGRU and LatentCross on both settings (i.e. fixed geographical distances $\Delta g = 1$ km and $\Delta g = 5$ km).

Moreover, Figure 6.3 demonstrates that the effectiveness of all approaches for the Brightkite dataset decreases as the time intervals between two successive checkins increase

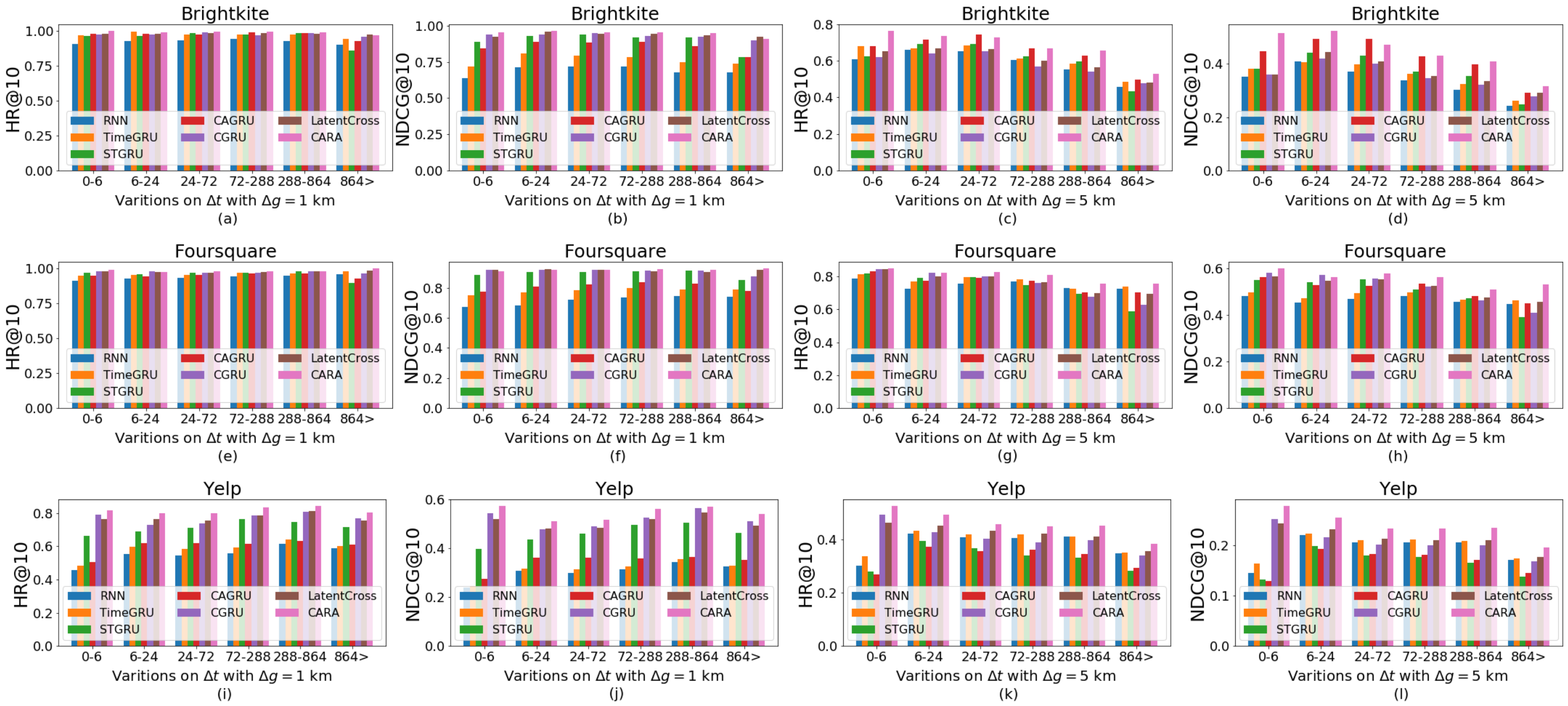


Figure 6.3: Performances between our proposed CARA architecture and various GRU architectures on the Brightkite, Foursquare and Yelp datasets by varying the time interval Δt in terms of hours with the fixed values of the geographical distances Δg (1 and 5 km).

because users are less likely to be influenced by venues they visited less recently. However, the experimental results from Figures 6.3 (a)-(b), (e)-(f) and (i)-(j) using a fixed geographical distance of $\Delta g = 1$ km on the Brightkite, Foursquare and Yelp datasets, respectively, demonstrate the ability of CARA in capturing the users' contextual dynamic preferences (as discussed in Section 6.1). In particular, even when the time interval between two checkins is long (e.g. more than 864 hours) but the geographical distances are small (i.e. $\Delta g = 1$ km), CARA still outperforms all baselines, demonstrating the value of learning using nearby checkins as well as using the recent checkins.

6.9 Conclusions

In this chapter, we proposed a novel Contextual Attention Recurrent Architecture (CARA) for Context-Aware Venue Recommendation (CAVR) that can effectively capture the users' contextual *dynamic* preferences from the users' sequence of checkins. In particular, the CARA architecture consists of two gating mechanisms: namely (1) the Contextual Attention Gate (CAG) that controls the influence of the ordinary context on the users' contextual *dynamic* preferences and (2) the Time-and Spatial-based Gates (TSG) that control the influence of the hidden state of the previous GRU units based on the time intervals and the geographical distances between two successive checkins. Together with the CAG and TSG gates, our proposed CARA architecture aims to address the three elicited limitations of the existing state-of-the-art GRU architectures: namely **Limitations G2-G4** of TimeGRU, CGRU and LatentCross. In particular, unlike the time gate proposed by (Zhu et al., 2017) (TimeGRU), our proposed TSG gates are sufficiently flexible to incorporate multiple types of transition contexts (e.g. the time interval and the geographical distance between two successive checkins). Moreover, unlike the previous state-of-the-art CGRU and LatentCross architectures that treat the ordinary and transition contexts equally, with the CAG and TSG gates, our proposed CARA architecture can effectively and independently control the impacts of ordinary and transition contexts that can influence the users' contextual *dynamic* preferences.

Overall, our comprehensive experiments on three large-scale datasets from the Brightkite, Foursquare and Yelp commercial LBSNs demonstrated the significant improvements of our proposed CARA architecture for CAVR in comparison with various previous state-of-the-art GRU architectures, as well as various recently proposed factorisation approaches, in both normal and cold-start settings. In particular, in answering research question RQ6.1, our experimental results showed that our proposed CARA architecture can effectively address **Limitation G2** of TimeGRU, in which DRCF can consistently and significantly outperform TimeGRU in terms of the HR and NDCG measures by 12.46% and 29.16%, respectively, on the average across the three used datasets (see Table 6.5). Next, by answering research questions RQ6.2 and RQ6.3, the experimental results demonstrated that CARA can address **Limitations G3-G4** of the existing state-of-the-art GRU architectures (CGRU and LatentCross). In particular, the experimental results showed that CARA can address **Limitation G3** of CGRU and LatentCross where CARA can significantly and consistently outperform CGRU and LatentCross across the three used datasets in terms of the HR and NDCG metrics by approximately 5%. These results show that the *transition* and *ordinary* contexts are different and should be treated separately (**Limitation G3**). Next, by answering research question RQ6.3, we can show that the absolute continuous values of the *transition* context used by CARA can preserve the influence of successive checkins more effectively than the quantised mapping values of the *transition* context used by the state-of-the-art CGRU and LatentCross architectures. Finally, by answering research question RQ6.4, we demonstrated that the TSG gates of the CARA architecture can effectively incorporate multiple types of *transition* context, thereby addressing **Limitation G2** of TimeGRU.

In our thesis statement (see Section 1.2), we hypothesised that the quality of context-aware venue recommendation can be effectively enhanced by leveraging the sequential order of the user's checkins and the contextual information associated with the successive checkins. To achieve this, an effective Recurrent Neural Network architecture, which captures the user's contextual short-term (*dynamic*) preferences from the sequence of user's checkins is needed. Based upon our experiments in this chapter, we conclude that our proposed CARA architecture, which can effectively capture the users' contextual *dynamic* preferences from the users' sequence of checkins can enhance the quality of context-aware venue recommendations. This far, we have conducted that the sequential order of users' checkins as well as the contextual information associated with the users' successive checkins play an important role in improving the quality of context-aware venue recommendations. Although our proposed CARA architecture can effectively capture the users' contextual *dynamic* preferences from the users' sequence of checkins, it still relies on the dot product of latent factors (see Equation (6.2)) to estimate the preference of a user for a given venue at a particular time context. However, previous work (He et al., 2017) and our experiments conducted in Section 5.8 demonstrated that the dot product operation is not sufficiently effective to capture the complex structure of user-venue interactions. In the last of our technical chapters, Chapter7, we will propose a Contextual Recurrent Collaborative Filtering framework that integrates the CARA architecture into the DRCF framework to effectively generate higher quality context-aware venue recommendations.

Chapter 7

Contextual Recurrent Collaborative Filtering Framework

7.1 Introduction

In the previous chapter, we described our proposed Contextual Attention Recurrent Architecture (CARA), an extension of the traditional Gated Recurrent Unit (GRU) architecture that effectively incorporates different types of contextual information associated with the users' sequence of checkins for Context-Aware Venue Recommendation (CAVR). Our experimental results in the previous chapter demonstrated the effectiveness of the CARA architecture in generating effective context-aware top-K venue recommendations in comparison with various state-of-the-art GRU architectures and factorisation approaches. Similar to the existing GRU architectures in the literature (Smirnova and Vasile, 2017; Beutel et al., 2018; Zhu et al., 2017), described in Section 3.6.2, the CARA architecture still relies on a dot product of latent factors of users and venues to capture the users' contextual *dynamic* preferences in a Collaborative Filtering manner (**Limitation C1**). However, previous work (He et al., 2017) and our experiments conducted in Section 5.8 demonstrated that the dot product operation is not sufficiently effective to capture the complex structure of user-venue interaction. **Limitation C1** of the CARA architecture can be summarised below:

Limitation C1: There is a disadvantage in the CARA architecture that still relies on the dot product operation to combine the latent factors of users and venues when predicting a user's checkin.

In Chapter 5, we described our proposed Deep Recurrent Collaborative Filtering framework (DRCF) that leverages the Multi-Layer Perceptron (see Section 2.2.1) and the traditional RNN models (see Section 2.2.3) to learn the complex structures of the users' se-

Table 7.1: Summary of the DRCF framework and the CARA architecture and their corresponding limitation

	DRCF	CARA	CRCF
Context-aware venue recommendation	×	✓	✓
Rely on different operations	✓	×	✓
Limitation	D1	C1	-

quences of checkins. In particular, instead of relying on the dot product of the latent factors, the DRCF framework relies on a neural architecture that can learn an arbitrary function from the sequences of user' checkins. However, the DRCF framework was initially proposed for venue recommendation and not suitable for CAVR because it still relies on the tradition RNN models, which are not sufficiently flexible to incorporate the user's preferred context as well as the contextual information associated with the user's sequences of checkins (**Limitation D1**). **Limitations D1** of the DRCF framework can be summarised as follows:

Limitation D1: There is a disadvantage in the DRCF framework that models the user's short-term (*dynamic*) preferences from sequential order of checkins by leveraging only the sequence of previously visited venues and ignoring the context associated with the checkins.

In this chapter, we aim to address **Limitation D1** of the DRCF framework and **Limitation C1** of the CARA architecture by proposing a novel Contextual Recurrent Collaborative Filtering framework (CRCF). The CRCF framework is an extension of the DRCF framework that leverages the sequence of the users' checkins, the users' preferred context and the contextual information associated with the sequence of the users' checkins to effectively and comprehensively capture the users' contextual long-term (*static*) and short-term (*dynamic*) preferences for CAVR. In particular, to effectively capture the users' contextual *dynamic* preferences from their sequence of checkins, we propose to integrate our proposed CARA architecture into the CRCF framework. Table 7.1 provides the summary of the CRCF framework, the DRCF framework and the CARA architecture and their corresponding limitation. Although, later in Section 7.4, we will demonstrate that the proposed CRCF framework can address the limitations of the DRCF framework and the CARA architecture, CRCF may introduce new limitations. For example, one possible limitation of the CRCF framework is that it is not sufficiently flexible to leverage the social information to improve the quality of context-aware venue recommendation. Overall, our contributions are summarised below:

- Section 7.2 provides an overview of our proposed Contextual Recurrent Collaborative Filtering framework (CRCF) for CAVR. In particular, we describe how to extend the DRCF framework to take the users' preferred context into account to and to integrate the CARA architecture into the CRCF framework to effectively generate high quality

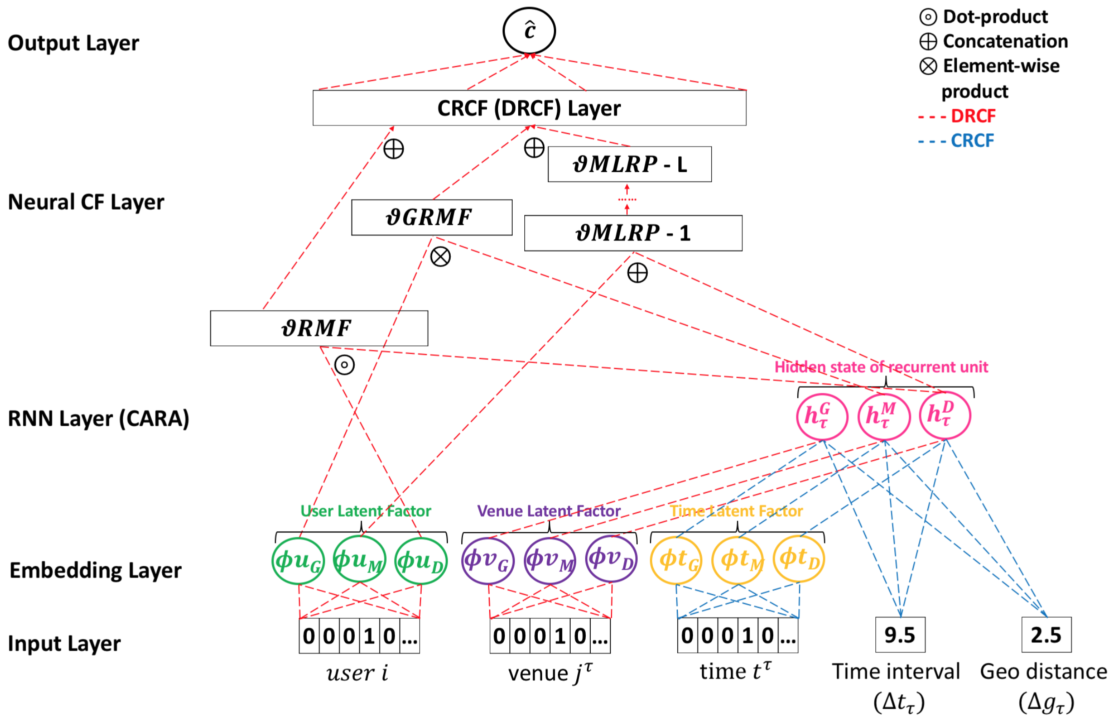


Figure 7.1: A diagram of Contextual Recurrent Collaborative Filtering Framework. The connections of each layer linked by the red-dashed lines illustrate the DRCF framework.

of context-aware venue recommendations.

- Section 7.3 presents our experimental methodology in terms of datasets and measures as well as algorithm parameters.
- In Section 7.4, we empirically evaluate the effectiveness of our proposed CRCF framework in comparison with the existing factorisation approaches as well as our proposed DRCF framework (see Chapter 5) and our proposed CARA architecture (see Chapter 6). Moreover, we investigate the robustness of the CRCF framework by leveraging risk analyses techniques proposed by Wang et al. (2012) and Dinçer et al. (2014).
- Section 7.5 provides a summary of this chapter.

7.2 Contextual Recurrent Collaborative Filtering Framework (CRCF)

In this section, we describe a Contextual Recurrent Collaborative Filtering framework (CRCF), an extension of the DRCF framework, that can effectively incorporate different types of contextual information associated with the sequential feedback (i.e. the time interval and geographical distance between two successive checkins) to model users' short-term (dynamic) preferences. In particular, the CRCF framework aims to generate a ranked-list of venues that a user might prefer to visit at time t based on the sequences of checkins $s_{i,t}$. The CRCF framework consists of five layers - the connections between these layers are presented using both blue- and red-dashed lines in Figure 7.1. The structure of the CRCF framework is different from the structure of the DRCF framework including its inputs, embedding and RNN layers because the CRCF can leverage the contextual information, whereas the DRCF framework cannot. Starting at the bottom of the figure, at the input layer, at time step τ, given a user i, venue j and time t^{τ}, we compute the time interval and the geographical distance between the given venue j and the venue k, which was previously visited at time step $\tau - 1$, as $\Delta t^{\tau} = t^{\tau} - t^{\tau-1}$ and $\Delta g_{\tau} = dist(lat_j, lng_j, lat_k, lng_k)$, respectively. $dist()$ is the Haversine distance function that returns the distance between the given latitudes and longitudes. In the embedding layer, there are three additional embedding layers highlighted in yellow in Figure 7.1 that are used to generate the latent factors of the time $\phi t^{\tau} \in \mathbb{R}^d$. Note that we only consider the time of checkins as the user's preferred context. However, our proposed framework is flexible to support multiple types of context (e.g. the current weather of the day). Next, the latent factors of venue and time (ϕv_j^{τ} and ϕt^{τ}) as well as the time interval Δt_{τ} and the geographical distance Δg_{τ} are fed into the RNN layer. In the RNN layer, we exploit the CARA architecture rather than the traditional RNN models used by the DRCF framework to encapsulate the *dynamic* user preferences. In particular, the main advantage of the CARA architecture over the traditional RNN models is that it can effectively capture the users' *dynamic* preferences by taking the contextual information associated with the users' two successive checkins into account. The output of the recurrent layer is the hidden state of the recurrent unit at time step τ, $h_{\tau} \in \mathcal{R}^d$, which is defined as follows:

$$h_{\tau} = f_{CARA}(\phi v_j^{\tau}, \phi t^{\tau}, \Delta t_{\tau}, \Delta g_{\tau}; \theta_r) \tag{7.1}$$

where $\theta_r = \{W, R, U, b\}$ denotes the set of parameters of the recurrent layer. Further details of the CARA architecture, f_{CARA}, are described in Section 6.2. Then, similar to the DRCF framework, the latent factors of user ϕu_i, and the user's *dynamic* preferences h_{τ} are fed into the Neural CF layer and the output layer, respectively. The objective function of the CRCF framework is similar to DRCF's, as described in Equation (5.2).

There are two advantages of the CRCF framework over either the DRCF framework or the CARA architecture. First, CRCF allows to take the user's context into account to generate effective venue recommendation based on his/her context, while DRCF cannot. Although CARA can incorporate the user' context during the recommendation process, it still relies on the dot product of the latent factors when making recommendations. Indeed, previous work (He et al., 2017) and our experiments in Chapter 6 have shown that the dot product operation is not effective in capturing the complex structure of user-venue interactions. Unlike CARA, our proposed CRCF framework is built upon the DRCF framework, which exploits the element-wise product and the concatenation operation to effectively capture the complex structure of the user-venue interactions.

7.3 Experimental Methodology

In the remaining of this chapter, we evaluate the effectiveness and robustness of our proposed Contextual Recurrent Collaborative Filtering (CRCF) framework in comparison with various matrix factorisation-based approaches. In particular, the CRCF framework aims to address **Limitation D1** of the CRCF framework (see Chapter 5) and **Limitation C1** of the CARA architecture (see Chapter 6). To address **Limitations D1 & C1**, we aim to answer the following research questions:

RQ7.1 *Can we enhance (a) the effectiveness and (b) the robustness of the Contextual Recurrent Collaborative Filtering (CRCF) framework for CAVR, by exploiting the state-of-the-art Contextual Attention Recurrent Architecture (CARA) to leverage the time interval and the geographical distance associated with sequences of checkins?*

RQ7.2 *Can the* dynamic *geo-based negative sampling approach that leverages both the sequential properties of checkins and the geographical location of venues, enhance (a) the effectiveness and (b) the robustness of CRCF and alleviate the cold-start problem?*

Table 7.2 summarises the research questions we aim to address in this chapter and their corresponding success decision. To demonstrate that our proposed CRCF framework can address **Limitation D1** of DRCF and **Limitation C1** of CARA, we aim to answer research questions RQ7.1(a) and RQ7.1(b) by comparing the performances of the CRCF framework with the DRCF framework and the CARA architecture as well as existing factorisation approaches. Next, by answering research question RQ7.2(a), we can demonstrate that our proposed *dynamic* geo-based negative sampling approach, described in Section 5.3, can enhance the effectiveness of the CRCF framework in generating context-aware top-K venue recommendations. Furthermore, we aim to answer research question RQ7.2(b) to demonstrate that the CRCF framework with the *dynamic* geo-based negative sampling approach can alleviate the cold-start user problem for Context-Aware Venue Recommendation (CAVR).

Table 7.2: Summary of each research question and its corresponding success decision.

Research Question	Limitation	Success Decision
RQ7.1(a)	D1, C1	The CRCF framework is more effective than the DRCF framework and the CARA architecture in generating context-aware top-K venue recommendations
RQ7.2(a)	-	The CRCF framework with our proposed *dynamic* geo-based negative sampling approach is more effective in alleviating the cold-start problem than the CRCF framework with the traditional negative sampling approach used in BPR
RQ7.1(b)	D1, C1	The CRCF framework is more robust (i.e. a likelihood to underperform by a given baseline model) than the DRCF framework and the CARA architecture
RQ7.2(b)	-	The CRCF framework with our proposed *dynamic* geo-based negative sampling approach is more robust in alleviating the cold-start problem than the CRCF framework with the traditional negative sampling approach used in BPR

In the remainder of this section, we describe the experimental setup in terms of datasets and measures (Section 7.3.1), baselines (Section 7.3.2), algorithm parameters (Section 7.3.3) and measures (Section 7.3.4). The experimental results and analysis follow in Section 7.4.

7.3.1 Datasets

Similar to the experimental setup used in Chapters 5 & 6, we conduct experiments using three publicly available large-scale user-venue interaction datasets from LBSNs. In particular, to show the generalisation of our proposed framework across multiple LBSN platforms and sources of feedback evidence, we use two checkin datasets from Brightkite[1] and Foursquare[2], and a rating dataset from Yelp[3]. Following the common practice from previous works (Rendle et al., 2009; He et al., 2017), we remove venues with less than 10 checkins. Table 7.3 summarises the statistics of the filtered datasets. To evaluate the effectiveness of our proposed CRCF framework, we adopt a *leave-one-out* evaluation methodology, previously used in our experiments in Chapter 5 and Chapter 6: for each user, we select his/her most recent checkin as a ground truth and randomly select another 100 venues that the user

[1] https://snap.stanford.edu/data/

[2] https://archive.org/details/201309_foursquare_dataset_umn

[3] https://www.yelp.com/dataset_challenge

Table 7.3: Statistics of the three used datasets.

	Brightkite	Foursquare	Yelp
Number of normal users	14,374	10,766	38,945
Number of venues	5,050	10,695	34,245
Number of ratings or checkins	681,024	1,336,278	981,379
Number of cold-start users	5,578	154	6903
% density of User-Venue matrix	0.93	1.16	0.07

has not visited before as the testing set, where the remaining checkins are used as the training and validation set. The context-aware venue recommendation task is thus to rank those 101 venues for each user, given his preferred context (i.e. time), aiming to rank the highest the most recent ground truth checkin. Note that the context-aware venue recommendation task allows to recommend venues that the user has previously visited, for example in a different context. For instance, while a user may have visited a restaurant a week ago, recommending the same restaurant to the user to visit in the next few hours is acceptable. Moreover, we conduct two separate experiments, namely: *Normal Users* (those with ≥ 10 checkins) and *Cold-start Users* (< 10 checkins) to evaluate the effectiveness of our proposed CRCF framework in the general and cold-start settings.

7.3.2 Baselines

We compare our proposed Contextual Recurrent Collaborative Filtering (CRCF) framework with various matrix factorisation-based approaches. We implement all baselines and the CRCF framework using Keras[4], a deep learning framework built on top of Theano[5]. Our baselines can be grouped into two different groups: namely MF- and RNN-based approaches. Our implementation of the CRCF framework is available as open source[6]. Note that some baselines may not be originally proposed for venue recommendation but are sufficiently flexible to be applied to such a task without any disadvantage. Similar to the experimental setup in Chapter 6, for a fair comparison, the choice of recurrent models for the RNN-based factorisation baselines is fixed to the Gated Recurrent Units (GRU) proposed by (Chung et al., 2014), described in Section 2.2.3.3. In addition, compared to LSTM, GRU has less parameters yet is as effective as LSTM for recommendations (Smirnova and Vasile, 2017; Tan et al., 2016; Tang et al., 2017). Note that we omit varying the choice of recurrent models (e.g. LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Chung et al., 2014)) and RNN settings, which have already been explored in the literature (e.g. (Tan et al., 2016; Tang et al., 2017)). Our baselines are summarised below:

[4] https://github.com/fchollet/keras

[5] http://deeplearning.net/software/theano

[6] https://github.com/feay1234/CRCF

MostPop is a baseline that ranks venues in descending order of the venues' popularities, calculated across all users.

MostVisit is a baseline that ranks venues for a given user in descending order of the venues' popularity for that user.

RecentVisit is a baseline that takes the user's sequential order of checkins into account and recommends the most recently visited venue to the user.

MF (Koren et al., 2009) is the traditional matrix factorisation approach that aims to accurately predict the users' checkin on the unvisited venues.

BPR (Rendle et al., 2009) is the classical pairwise ranking approach, coupled with matrix factorisation for user-venue checkin prediction (see Section 2.1.1.1).

GeoBPR (Yuan et al., 2016) is an extension of BPR that incorporates the geographical location of venues to sample negative venues that are far away from the user's previous visits (see Section 3.5.1).

RNN-MF (Zhang et al., 2014b) is a sequential click prediction with recurrent neural networks approach (see Section 3.6.1).

DREAM (Yu et al., 2016) is a RNN model that incorporates BPR for ranking optimisation. As DREAM is originally proposed for next shopping-basket recommendation, to permit a fair comparison with our proposed CRCF framework, we reimplement DREAM to treat a single checkin as the shopping-basket purchase (see Section 3.6.1).

NeuMF (He et al., 2017) is the Neural Matrix Factorisation framework[7], which exploits Multi-Level Perceptron (MLP) models to capture the complex structure of user-item interactions (see Section 3.4.2).

DRCF is our proposed Deep Recurrent Collaborative Filtering framework for venue recommendation that extends the NeuMF framework to exploit the RNN-based models to model the sequences of the users' checkins (see Section 5.2).

STELLAR (Zhao et al., 2016) is a Spatial-TEmporaL LAtent Ranking framework for CAVR that aims to recommend the list of venues based on the user's preferred time and last successive visits. Note that this is the only context-aware framework that does not rely on the RNN-based approaches to model the users' sequential order of checkins.

CARA is our proposed Contextual Attention Recurrent Architecture[8] for CAVR that leverages the contextual information associated with the sequence of user's checkins to model the users' contextual *dynamic* preferences (see Section 6.2).

[7] https://github.com/hexiangnan/neural_collaborative_filtering

[8] https://github.com/feay1234/CARA

7.3.3 Recommendation Parameter Setup

Similar to the parameter setup in Chapter 5 and Chapter 6, we follow (He et al., 2017; Zhao et al., 2016; Yu et al., 2016) to set the dimension of the latent factors d and hidden layers h_τ of our proposed CARA architecture and all of the matrix factorisation-based approaches to be identical: $d = 10$ across three datasets. Furthermore, we randomly initialise all embeddings and recurrent layers' parameters, $\theta_r, \theta_e, \theta_h$, with a Gaussian distribution (with a mean of 0 and a standard deviation of 0.01) and apply the mini-batch Adam optimiser (Kingma and Ba, 2014) to optimise those parameters, which yields a faster convergence than SGD and automatically adjusts the learning rate for each iteration. We initially set the learning rate to 0.001[9] and set the batch size to 256. Since the impact of the recurrent parameters such as the size of the hidden state, have been explored in previous works (He et al., 2017, 2016b; Tan et al., 2016), we omit varying the size of the hidden layers and the dimension of the latent factors in this work. Indeed, larger sizes of hidden layers and dimensions may cause overfitting and degrade the generalisation of the models (He et al., 2016b, 2017; Tan et al., 2016)[10].

7.3.4 Measures

We measure the quality of the ranked list of venues in terms of Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG), which are similar to the measures used in Chapter 5 and Chapter 6. In particular, HR considers the ranking nature of the task, by taking into account the rank(s) of the venues that each user has previously visited/rated in the produced ranking, while NDCG goes further by considering the checkin frequency/rating value of the user as the graded relevance label. Finally, as before, significance tests are conducted using a paired t-test.

Furthermore, we experiment to determine the *robustness* of the CRCF framework, to measure its likelihood to underperform. To this end, we use risk-sensitive evaluation measures to quantify any underperformance in comparison to an established baseline recommender system. Throughout our robustness experiments, we use the Bayesian Personalised Ranking (BPR) model, which we argue that BPR is equivalent to BM25 baseline in web search, as the established baseline for venue recommendation system to evaluate the robustness of our proposed CRCF framework. To this end, we use risk-sensitive evaluation measures to quantify any underperformance compared to a given baseline model (i.e. the BPR model). All risk-sensitive measures are defined in terms of Risk & Reward (Wang

[9] The default learning rate setting of the Adam optimiser in Keras.

[10] The journal version of this chapter confirms that by increasing the size of hidden layers and dimension of the latent factors cause overfitting and degrade the generalisation of the models.

et al., 2012), where Risk is defined as the average reduction in effectiveness due to the use of the new target model in comparison to the baseline CF ranking model. In contrast, Reward is the positive improvement in effectiveness of the target model over the baseline model, averaged across all users. We use NDCG as the primary effectiveness measure for comparing the effectiveness of the new target model and the baseline CF ranking model. In particular, given a baseline CF ranking model (i.e. BPR), the Risk and Reward scores of using a target model (e.g. DRCF or CRCF) over the set of all users are measured as follows:

$$Reward = \frac{1}{|\mathcal{U}|}\sum_{i \in \mathcal{U}} max(0, M_t(i) - M_b(i)) \tag{7.2}$$

$$Risk = \frac{1}{|\mathcal{U}|}\sum_{i \in \mathcal{U}} max(0, M_b(i) - M_t(i)) \tag{7.3}$$

where M_b and M_t denote the effectiveness of the baseline CF ranking model and the target model for a given user i, respectively, calculated using NDCG. Let the overall gain of a target model be $Gain = reward - risk$. Next, a single measure that takes the risk-reward tradeoff into account is calculated as $U_{risk} = Gain - \alpha \cdot Risk$, where $\alpha \geq 0$ is a risk-sensitivity parameter (Wang et al., 2012). Note that with $\alpha = 0$, U_{risk} simply measures the average difference in performances between the two models across all users. On the other hand, increasing $\alpha > 0$ places more emphasis on penalising models that underperform compared to the baseline.

Following Dinçer et al. (2014), for $\alpha \geq 0$, a t-statistic can be formulated based on U$_{risk}$, which they called T$_{risk}$, and can be expressed as follows:

$$T_{risk} = \frac{U_{risk}}{SE(U_{risk})} \tag{7.4}$$

where $SE()$ is the standard error of the paired sample mean. The advantage of T$_{risk}$ over U$_{risk}$ is that it is easily interpreted for an inferential analysis of risk (i.e. if the system exhibits a significant level of risk for a given α). Indeed, T$_{risk}$ < 2 denotes a significant risk (Dinçer et al., 2014). Later in Section 7.4.2, we test the significance of an observed risk-reward tradeoff score between a target model and a given baseline by using T$_{risk}$ as the test statistic of the Student's t-test for matched pairs.

7.4 Experimental Results

In this section, we report the effectiveness and robustness of our proposed CRCF framework in comparison with various state-of-the-art factorisation approaches. In particular, to address research questions RQ7.1(a) and RQ7.2(a), we conduct various experiments to evaluate the effectiveness of the CRCF framework under the *Normal* and *Cold-Start* settings, which are discussed in Section 7.4.1. Moreover, to answer research questions RQ7.1(b) and RQ7.2(b), we further perform several risk analysis to investigate the robustness of the CRCF framework, which are discussed in Section 7.4.2

7.4.1 Effectiveness Evaluation

In this section, we report the effectiveness of our proposed CRCF framework in comparison with various state-of-the-art approaches. In particular, to answer research question RQ7.1(a), Section 7.4.1.1 reports the performance of the CRCF framework and the used baselines under the *Normal* and *Cold-Start* settings. In addition, to answer research question RQ2(a), Section 7.4.1.2 demonstrates the usefulness of the *dynamic* geo-based negative sampling approach in enhancing the effectiveness of the CRCF framework and alleviating the cold-start problem.

7.4.1.1 Effectiveness of the CRCF framework

Table 7.4 reports the effectiveness of the CRCF framework in comparison with various matrix factorisation-based approaches in term of the HR@10 and NDCG@10 measures on the three used datasets. In particular, the table contains two groups of rows, which report the effectiveness of various approaches under the *Normal Users* and *Cold-Start Users* experiments, respectively. Firstly, on inspection of the first group of rows of Table 7.4, we note that the relative venue recommendation quality of the baselines on the three datasets in terms of the two measures are consistent with the results reported for the various baselines in the corresponding literature (He et al., 2017; Yu et al., 2016; Zhang et al., 2014b). For instance, DRCF outperforms MF, BPR and NeuMF across the three datasets. Similarly, CARA outperforms STELLAR across the three datasets. Note that previous works (He et al., 2017; Yu et al., 2016; Zhang et al., 2014b) used different datasets, but our reimplementations of their proposed approaches obtain similar relative improvements.

Comparing CRCF with the various baselines, we observe that CRCF consistently and significantly outperforms all baselines for both HR and NDCG, across all datasets. In particular, comparing with DRCF and CARA, CRCF obtains 4.61%, 3-4.02% and 6.32-17.41% improvements in terms of NDCG for Brightkite, Foursquare and Yelp datasets, respectively. These results suggest that our proposed framework, an extension of DRCF that exploits the CARA architecture instead of the traditional RNN models to leverage the user's preferred context (i.e. time) and the contextual information associated with the sequence of checkins is more effective than the DRCF framework, which ignores those contexts. Moreover, comparing CRCF with the CARA architecture, which both take the users' context into account, the results imply that the neural architecture in the CRCF framework (i.e. an element-wise product and concatenation between the latent factors (see Figure 7.1)) can enhance the quality of venue recommendations. Such observation are consistent with the results reported by He et al. (2017) and our experimental results in Chapter 5.

Next, we note that unlike the Brightkite and Foursquare checkin datasets, the Yelp dataset consists only of user-venue ratings, and hence the sequential properties of visits to venues are less likely to be observed. We observe that the RNN-based approaches (RNN and DREAM) that take the sequential properties of checkins into account are more effective than the traditional MF-based approaches (MF and BPR) across the Brightkite and Foursquare checkin datasets. However, both RNN and DREAM are less effective than BPR for the Yelp rating dataset because the sequential properties of rating data are less pronounced than the other LBSNs. This is likely due to users writing Yelp reviews after visiting the venue. In contrast, our proposed CRCF framework is still the most effective across the different types of datasets, which is indicative of the generalisability of CRCF. In addition, we observe that CARA, which incorporates the contextual information, is as effective as DRCF on the two checkin datasets in terms of the two used measures[11], while CARA outperforms DRCF on the Yelp dataset. These results demonstrate that contextual information plays an important role in enhancing the effectiveness of CAVR. By integrating CARA into CRCF, we can further enhance the quality of CAVR across three datasets in terms of HR@10 and NDCG@10.

Within the second group of rows in Table 7.4, we further investigate the effectiveness of the CRCF framework by comparing with the baselines in the *Cold-Start Users* experiment. The results demonstrate that CRCF consistently and significantly outperforms all baselines across Brightkite and Yelp datasets on both measures. In particular, comparing the effectiveness of CRCF for cold-start users with DRCF and CARA, CRCF obtains 5.81-10% and 3.69-15.49% improvements in terms of NDCG, for the Brightkite and Yelp datasets, respectively. Although the performance of CRCF in alleviating the cold-start user problem is statistically indistinguishable from CARA and DRCF in the Foursquare dataset in terms of HR@10, CRCF significantly outperforms CARA in terms of NDCG@10 by 7%. This result suggests that the element-wise product and the concatenation of the latent factors used by CRCF play a more important role than the dot product of the latent factors used by CARA in

[11] Note that DRCF consists of three components (namely GRMF, MLRP and RMF models (see Section 5.2)), with each component having its own recurrent layer. Although CARA consists only one recurrent layer, it is as effective as DRCF. Moreover, in Section 6.6, we demonstrated that CARA significantly outperforms each individual component of DRCF.

generating more effective top-K venue recommendations for cold-start users. Furthermore, the results reported in the second group of rows in Table 7.4 demonstrate that our proposed CRCF framework is more effective than the DRCF framework and the CARA architecture in alleviating the cold-start user problem. Overall, in response to research question RQ7.1(a), we find that our proposed CRCF framework, which leverages the sequences of the users' checkins as well as the contexts associated with the checkins, is effective for CAVR for both the *Normal* and *Cold-Start* users.

7.4.1.2 Usefulness of *Dynamic* Geo-based Negative Sampling

In this section, to address research question RQ7.2(a), we evaluate the usefulness of the *dynamic* geo-based negative sampling approach, denoted with the suffix $_{dgeo}$) in enhancing the robustness of our proposed CRCF framework that used the traditional BPR negative sampling approach. Similar to Table 7.4, Table 7.5 reports the observed performances of the CRCF and DRCF frameworks as well as CARA architecture when incorporating our proposed *dynamic* geo-based negative sampling approach, described in Section 5.3, which takes the geographical location of venues into account during the negative sampling process.

In the first group of rows in Table 7.5, we analyse the effectiveness of our proposed CRCF framework by comparing with the state-of-the-art DRCF framework and CARA architecture when incorporating our proposed *dynamic* geo-based negative sampling approach, described in Section 5.3, in the *Normal Users* experiment. First, we observe similar results to those reported in Section 5.8, namely that the negative sampling approach can significantly improve the effectiveness of DRCF, CARA and CRCF in terms of HR@10 and NDCG@10 across the three datasets. For example, CRCF$_{dgeo}$ obtains over 6.65%, 3.1% and 11.72% improvements over CRCF in terms of HR@10 on the Brightkite, Foursquare and Yelp datasets, respectively. Note that DRCF$_{dgeo}$ and CARA$_{dgeo}$ also obtain similar percentage improvements over DRCF and CARA, respectively, across the three datasets. In addition, CRCF$_{dgeo}$ consistently and significantly outperforms all baselines that consider the geographical location of venues during the negative sampling process (i.e. GeoBPR, DRCF$_{dgeo}$ and CARA$_{dgeo}$) across all three datasets. These improvements and observed results demonstrate that the *dynamic* geo-based negative sampling approach plays a crucial role in enhancing the effectiveness of DNN-based approaches. In addition, Figure 7.2 reports the test performances of CRCF$_{dgeo}$ and the baselines for each of the three datasets with all users over each training iteration. From the figure, we observe that CRCF$_{dgeo}$ outperforms all the baselines at every iteration and converges faster than others across the three datasets. Moreover, we observe that both DRCF$_{dgeo}$ and CARA$_{dgeo}$ are more effective that DRCF and CARA. However, on the Yelp dataset, we find that CRCF, which relies on the traditional BPR negative sampling approach (Rendle et al., 2009), is more effective than DRCF$_{dgeo}$ at every iteration in terms of HR@10 and NDCG@10. These results demonstrate that the users' context plays an important role in enhancing the quality of CAVR. Indeed, the *dynamic* geo-based negative sampling approach may not be useful when the sequential properties of the users' observed feedback are less likely to be observed, as in the Yelp rating dataset. Hence, DRCF$_{dgeo}$ is less effective than CRCF for both measures on the Yelp dataset.

Next, within the second group of rows in Table 7.5, we further investigate the effectiveness of CRCF$_{dgeo}$, DRCF$_{dgeo}$ and CARA$_{dgeo}$, which all rely on the *dynamic* geo-based negative sampling approach, in the *Cold-Start Users* experiments. First, similar to the *Normal Users* experiments, we observe that the dynamic geo-based negative sampling approach can significantly improve the effectiveness of DRCF, CARA and CRCF in terms of HR@10 and NDCG@10 across the three datasets in the *Cold-Start Users* experiments. In particular, the results demonstrate that CRCF$_{dgeo}$ consistently and significantly outperforms all baselines across both the Brightkite and Yelp datasets on both measures. In particular, comparing the effectiveness of alleviating the cold-start users of CRCF$_{dgeo}$ with DRCF$_{dgeo}$ and CARA$_{dgeo}$, CRCF obtains approximately 5$\%$, 4.42 - 12.5$\%$ improvements in terms of HR@10 for the Brightkite and Yelp datasets, respectively. Although the effectiveness of CRCF$_{dgeo}$ for the cold-start users is less than that of DRCF$_{dgeo}$ and CARA$_{dgeo}$ for the Foursquare dataset, there is no significant difference between CRCF$_{dgeo}$, DRCF$_{dgeo}$ and CARA$_{dgeo}$ in terms of HR and NDCG on the Foursquare dataset. These results demonstrate that the *dynamic* geo-based negative sampling approach can enhance the effectiveness of CRCF, DRCF and CARA in generating effective CAVR for the cold-start users.

We further investigate the usefulness of the dynamic geo-based negative sampling approach and the CARA architecture in enhancing the effectiveness of CRCF$_{dgeo}$ under different settings for CAVR. Note that CARA leverages the time interval and the geographical distance between two successive checkins to model the user's dynamic preferences, motivating the integration of CARA into our proposed CRCF framework (as described in Section 7.3). In particular, the plots in Figure 7.3 present the performances on the Brightkite, Foursquare and Yelp datasets – in terms of NDCG@10 – of various approaches, by considering the users with particular time intervals Δt (hours) and geographical distances Δg (km) between their last checkin and ground-truth checkin. For example, if a user checks in at venues A, B and C in a sequence, his/her last checkin is at venue B and his/her ground-truth checkin is at venue C. Then, we calculate the distance Δg and the time interval Δt between venues B and C. Note that these two checkins may occur at the same venue, hence the distance $\Delta g = 0$, while the time interval Δt between these two checkins is such that $\Delta t > 0$.

First, the results from Figure 7.3 demonstrate that CRCF consistently outperforms

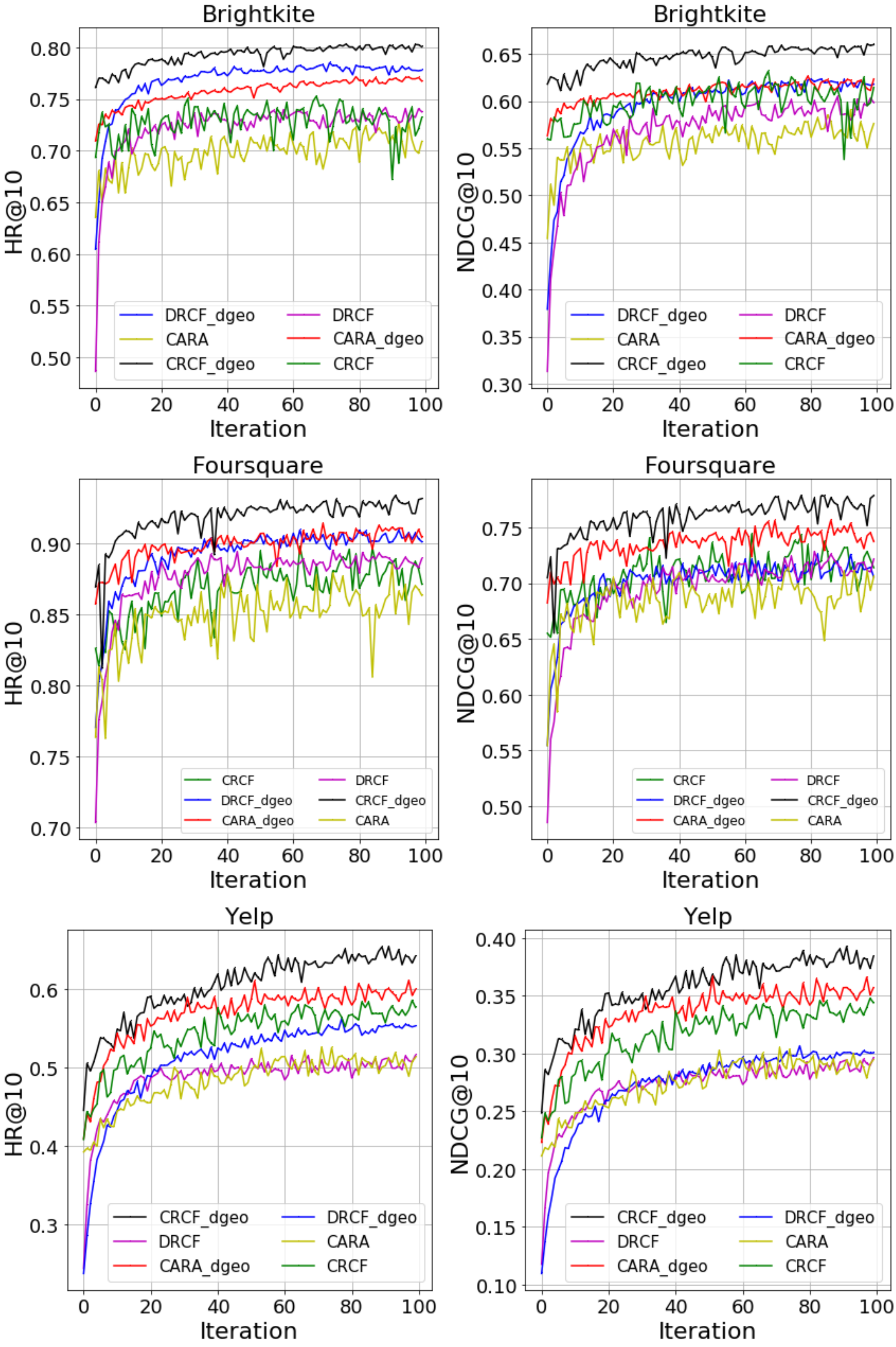


Figure 7.2: Test recommendation performances in terms of HR & NDCG of various approaches with respect to the number of iterations.

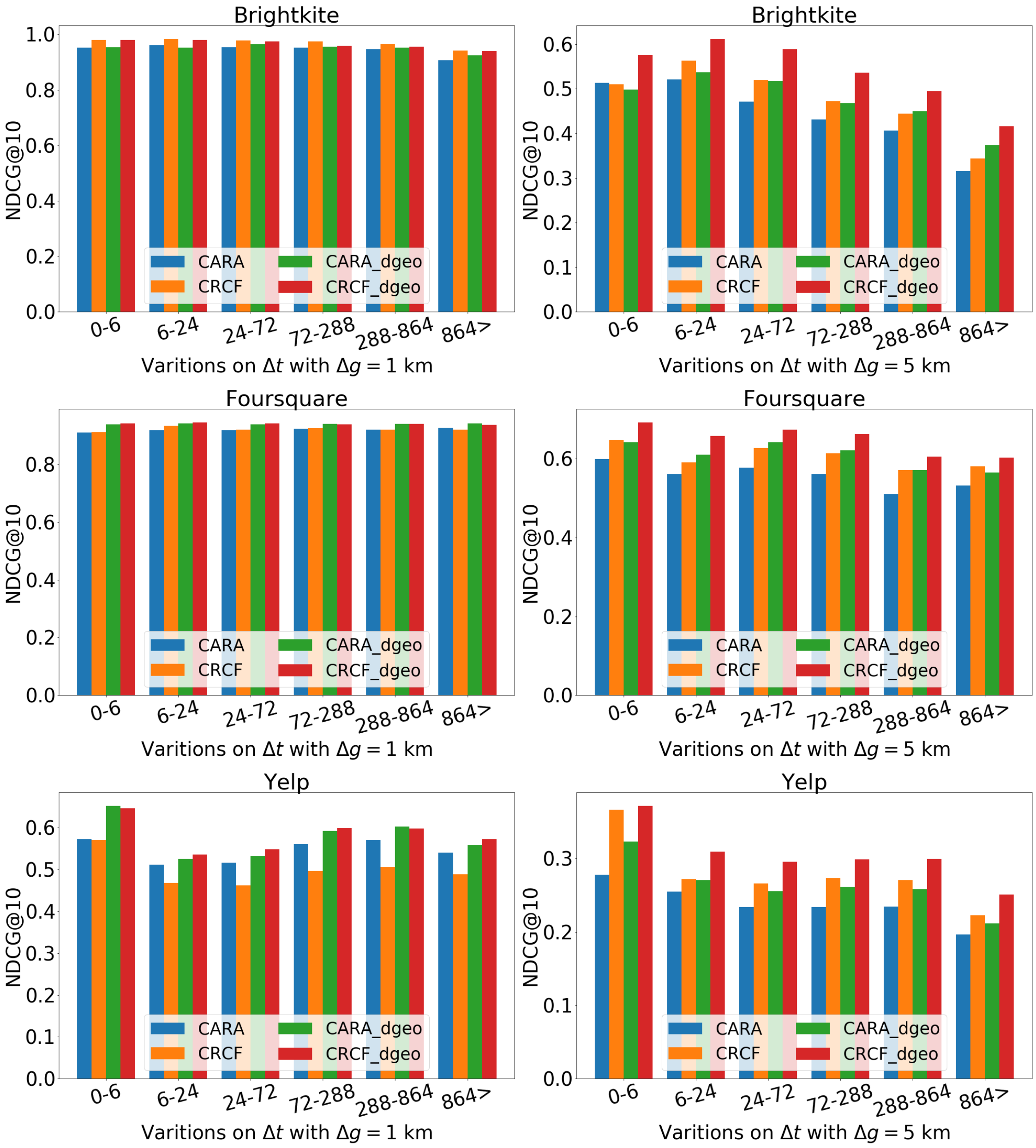


Figure 7.3: Performances of various approaches in terms of NDCG@10 on the Brightkite, Foursquare and Yelp datasets by varying the time interval Δt in terms of hours with the fixed values of the geographical distances Δg (1 and 5 km).

CARA across the three datasets in terms of NDCG@10 on various time intervals Δt and geographical distances Δg. These results suggest that the neural architecture (i.e. the element-wise and concatenation operations of latent factors, described in Section 7.3) in CRCF can effectively integrate CARA, hence obtaining the improvements over CARA on both settings. Moreover, the experimental results using a fixed geographical distance of $\Delta g = 5$ km on the right hand plots in Figure 7.3 demonstrate that the effectiveness of all approaches on the three datasets decreases as the time intervals between two successive checkins increases. These observations suggest that users are less likely to be influenced by distant venues they visited a long time ago, which are consistent with results previously reported in Section 6.6. In contrast, the performances of all approaches on a fixed geographical distance of $\Delta g = 1$ km setting are relatively stable on the Brightkite and Foursquare datasets. Intuitively, nearby venues visited by users are more likely to influence the users' preferences for their next venues regardless of when those nearby venues were visited. As mentioned above, the sequential properties are less likely to be observed from the user-venue rating Yelp dataset. Hence, unlike the Brightkite and Foursquare checkin datasets, the *dynamic* geo-based negative sampling approach may not be useful in enhancing the performances of DRCF, CARA and CRCF on the Yelp dataset. Furthermore, comparing the approaches that apply the dynamic geo-based negative sampling approach, we find that the effectiveness of both CARA$_{dgeo}$ and CRCF$_{dgeo}$ across the three datasets on different settings can be enhanced by the dynamic negative sampling approach. In particular, CRCF$_{dgeo}$ is the most effective approach compared to all baselines across the three datasets on various settings. Overall, in response to research question RQ7.2(a), we find that the dynamic geo-based negative sampling approach can effectively improve the performances of the CRCF framework for CAVR on various settings that consider different time intervals and geographical distances between user's two successive checkins.

7.4.2 Robustness Evaluation

In this section, we evaluate the robustness of the CRCF and DRCF frameworks as well as the CARA architecture using the risk-sensitive measures (i.e. Reward & Risk and U$_{risk}$), proposed by Wang et al. (2012) to quantify any underperformance of DRCF, CARA and CRCF compared to the BPR model (Section 7.4.2.1).[12] Apart from the risk-sensitive measures, we also use the T$_{risk}$ measure, proposed by Dinçer et al. (2014), to evaluate whether a given framework or model exhibits a significant risk compared to the BPR model. In particular, we test the significance of an observed risk-reward tradeoff score between a target model and

[12] Indeed, we argue that BPR is a widely used baseline in recommendation systems, which is akin to the use of BM25 in web search, and hence is appropriate as our robust baseline for risk-sensitive evaluation.

the BPR model by using T_{risk} as the test statistic of the Student's t-test for matched pairs. In addition, we evaluate the usefulness of the *dynamic* geo-based negative sampling approach in enhancing the robustness of the CRCF framework (Section 7.4.2.2).

7.4.2.1 Robustness of the CRCF framework

Tables 7.6 & 7.7 report the robustness of the CRCF framework in comparison with the DRCF framework and CARA architecture on the three datasets in terms of different measures under the *Normal Users* and *Cold-Start Users* experiments, respectively. For instance, the Wins/Losses row shows the ratio of the number of users that benefit or do not benefit from a particular model compared to the BPR model. The lower the better for the Losses and Risk measures, while the higher the better for the Wins and Reward measures. On analysing Table 7.6, regarding the robustness of several approaches that do not apply the dynamic geo-based negative sampling approach (i.e. DRCF, CARA and CRCF), we find that CRCF is the most robust framework by consistently having the lowest Risk/Losses and the highest Reward/Wins in comparison with DRCF and CARA across the three datasets. In particular, CRCF can generate a more effective ranked list of venues than BPR (i.e. NDCG@10 is improved by CRCF compared to BPR) for 45.48%, 54.58% and 33.33% of users on the Brightkite, Foursquare and Yelp datasets, respectively. CRCF performs less effectively than BPR (i.e. NDCG@10 is degraded by CRCF compared to BPR) for 10.32%, 6.60% and 18.14% of users on the Brightkite, Foursquare and Yelp datasets, respectively. In addition, Figure 7.4 reports the wins-losses histogram of CRCF and the baselines on the three datasets. From the figure on the *Normal Users* experiments, we observe that CRCF consistently has consistently smaller changes in NDCG@10 on all bins on the left side (negative) of the vertical line and larger changes in NDCG@10 on the right side (positive) of the vertical line than the baselines across the three datasets. Moreover, we observe that, at $\alpha = 1$ (which emphasises risk twice over reward), the calculated U_{risk} scores of DRCF, CARA and CRCF are significantly higher than BPR, at $p < 0.05$ (as $\text{T}_{risk} > 2$), across the Brightkite and Foursquare datasets, while only the U_{risk} score of CRCF on the Yelp dataset exhibits significant risk ($\text{T}_{risk} < 2$). These results demonstrate that it is highly likely that CRCF will not perform worse than the BPR baseline across the three datasets, while both DRCF and CARA have $\text{T}_{risk} < -2$ at $\alpha = 1$ may underperform on the Yelp dataset (i.e. perform worse than BPR).

Next, Table 7.7 reports the risk measures for the *Cold-Start Users*, using the same notation as Table 7.6. In Table 7.7, we observe that CRCF consistently has consistently lower Risk/Losses and higher Reward/Wins than DRCF and CARA across the Brightkite and Yelp datasets for the *Cold-Start Users*. For example, CRCF is more robust than DRCF and CARA in terms of Wins as it can generates more effective venue suggestions than BPR for 2,394 users on the Brightkite dataset, while DRCF and CARA can only generate more effective venue suggestions than BPR for 2,183 and 2,221 users, respectively. Similar results in terms of Wins for DRCF, CARA and CRCF can also be observed on the Yelp dataset. In addition, CRCF exhibits less risk than DRCF and CARA at generating less effective venue suggestions than BPR. For example, on the Brightkite dataset, CRCF only generates less effective venue suggestions than BPR for 661 users, while DRCF and CARA generate less effective venue suggestions than BPR for 1,043 and 803 users, respectively. We observe similar results for DRCF, CARA and CRCF in terms of Losses on the Yelp dataset.

Moreover, we observe that, at $\alpha = 1$, the U$_{risk}$ scores of DRCF, CARA and CRCF denote significant improvements across the Brightkite and Foursquare datasets. These results demonstrate that there is no significant risk that these three approaches will perform worse than BPR baseline for the *Cold-Start* users for the Brightkite and Foursquare datasets. However, these three approaches are likely to be under the real risk on the Yelp dataset for both $\alpha = 1$ and $\alpha = 5$. Figure 7.4 shows that CRCF consistently has larger number of positive changes over BPR than DRCF and CARA on Brightkite and Yelp datasets across all positive bins, and likewise less for negative bins. However, for the Foursquare dataset, CRCF has higher Reward and lower Risk than DRCF on the average, while the number of Losses of DRCF is lower than CRCF. Likewise, the number of Wins of DRCF is higher than CRCF. For example, there is only 13% of cold-start users (20 out of 154 cold-start users) on Foursquare dataset whose recommendations generated by DRCF are less effective than BPR, while 52% of cold-start users (81 out of 154 cold-start users) on Foursquare suffer from CRCF's recommendations. These results can be clearly observed in Figure 7.4 on the Foursquare dataset for the *Cold-Start Users* experiments where CARA and CRCF obtain large number of negative changes in terms of NDCG@10 over BPR for $0.1 <$ NDCG 0.2. Overall, in response to research question RQ7.1(b), we find that our proposed CRCF framework is robust and less likely to perform worse than BPR baseline for CAVR for the *Normal* and *Cold-Start* users.

We further investigate the robustness of the CRCF framework in comparison with the DRCF framework and CARA architecture using T$_{risk}$ score. Figure 7.5 demonstrates the change in the T$_{risk}$ scores of the various approaches for various risk-sensitivity α parameter values from 0 to 15 under *Normal* and *Cold-Start Users* experiments. Note that, as mentioned in Section 7.3.1, the risk-sensitivity α parameter controls the risk-reward tradeoff of the U$_{risk}$ and T$_{risk}$. Indeed, as α increases, the tradeoff between risk and reward for each model changes in favour of risk compared to reward. T$_{risk}$ scores greater than $+2$ (indicated by red horizontal line in the figure) or less than -2 (indicated by blue-dashed horizontal line in

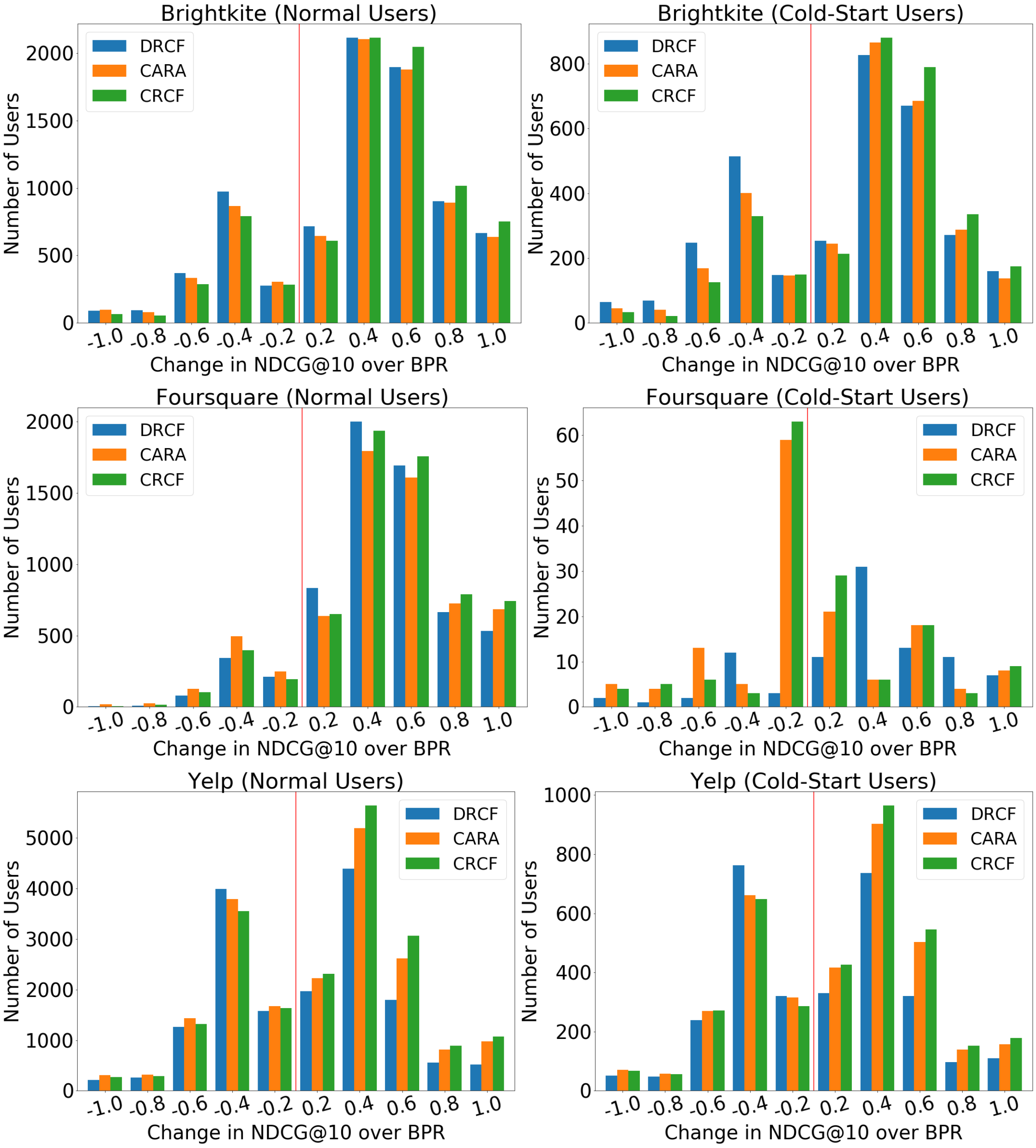


Figure 7.4: Typical Wins-Losses histogram of target new models (DRCF, CARA and CRCF) in comparison with the CF ranking baseline model, BPR, under the *Normal* and *Cold-start Users* settings. The vertical line in the figures separates wins from losses. We omit to report the number of users with no change in NDCG@10 over BPR (i.e. a target model is as effective as BPR).

the figure) exhibit significant differences from the baseline according to a two-tailed paired t-test with $p < 0.05$. On analysing the left hand figures in Figure 7.5, with respect to the *Normal* experiments, at $\alpha = 1$, we observe that all approaches (DRCF (BPR), CARA (BPR) and CRCF (BPR)) are significantly less risky than BPR across the three datasets. Moreover, we observe that as α increases, on the Brightkite and Yelp dataset, CRCF is significantly less risky than BPR when $\alpha = 4$ and $\alpha = 1$, respectively, while DRCF and CARA are not. On the Foursquare dataset, CRCF and DRCF are significantly less risky than BPR until $\alpha = 11$, while CARA is significantly less risky than BPR until $\alpha = 7$.

Next, on analysing the right hand plots in Figure 7.5, regarding the robustness of DRCF, CARA and CRCF under the *Cold-Start* experiments, we observe that, at $\alpha = 1$, all approaches are significantly less risky than BPR. However, as α increases to 3, CRCF is the only one approach that is less risky than the BPR baseline across the Brightkite and Foursquare datasets. Moreover, comparing CRCF with either DRCF or CARA, we observe that CRCF is only significantly less risky than DRCF when $\alpha = 1$ on the Brightkite dataset for the *Cold-Start* users. To further respond to research question RQ7.1(b), we find that our proposed CRCF framework is less risky for deployment to users, in that it only exhibits real risk compared to BPR for higher values of α than the existing state-of-the-art, DRCF and CARA, for both *Normal* and *Cold-Start* users experiments.

7.4.2.2 Usefulness of Dynamic Geo-based Negative Sampling for Robustness

In this section, in addressing research question RQ7.2(b), we evaluate the usefulness of the *dynamic* geo-based negative sampling in improving the robustness of the CRCF framework for the *Normal Users* experiments. In Table 7.8, we first observe that the *dynamic* geo-based negative sampling approach can consistently enhance the robustness of the CRCF framework across the three datasets. In particular, in comparison with DRCF$_{dgeo}$ and CARA$_{dgeo}$, CRCF$_{dgeo}$ is the most robust framework, as it generates more effective venue suggestions than BPR for 49.94%, 59.16% and 40.55% of users on the Brightkite, Foursquare and Yelp datasets, respectively. Moreover, comparing CRCF$_{dgeo}$ and CRCF, we observe that the dynamic geo-based negative sampling approach can enhance the Reward score of CRCF by approximately 4-7% and can reduce the Risk score of CRCF by approximately 0.6-3%. In addition, comparing the T$_{risk}$ scores of CRCF and CRCF$_{dgeo}$ on the Brightkite dataset, at $\alpha = 5$, we observe that CRCF$_{dgeo}$ is less likely to exhibit a real risk of performing worse than the BPR baseline, while CRCF is not. In addition, Figure 7.6 reports the robustness of CRCF and CRCF$_{dgeo}$ on the three datasets. From the left figures in Figure 7.6 on the *Normal Users* experiments, we observe that CRCF$_{dgeo}$ has consistently lower changes in NDCG@10 on all negative bins (i.e. the left side of the vertical line) and higher changes in NDCG@10 on all

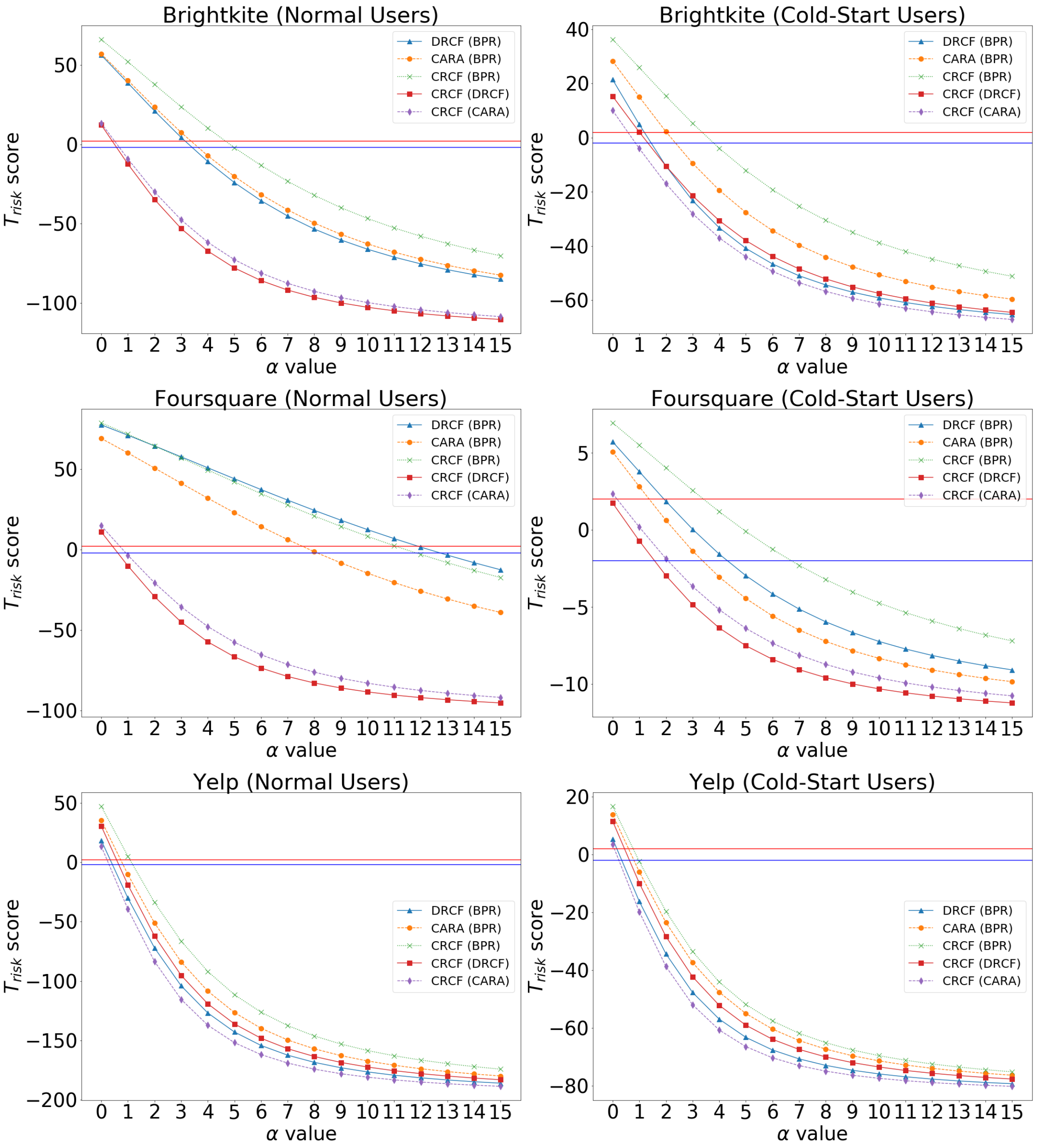


Figure 7.5: The changes in standardised T_{risk} scores for DRCF, CARA and CRCF with respect to the baseline, denoted inside the parentheses, over different α values under the *Normal* and *Cold-Start Users* experiments.

positive bin (i.e. the right side of the vertical line) than CRCF across the three datasets. Furthermore, Figure 7.7 reports wins-losses histograms of DRCF$_{dgeo}$, CARA$_{dgeo}$ and CRCF$_{dgeo}$ on the three datasets. From the left figures in Figure 7.7, on the *Normal Users* experiment, we observe that CRCF$_{dgeo}$ consistently has lower changes in NDCG@10 on all negative bins (i.e. the left side of the vertical line) and higher changes in NDCG@10 on all positive bins (i.e. the right side of the vertical line) than DRCF$_{dgeo}$ and CARA$_{dgeo}$ across the three datasets.

Next, we evaluate the usefulness of the *dynamic* geo-based negative sampling approach in improving the robustness of the CRCF framework for the *Cold-Start Users* experiments. Similar to the results reported in Table 7.8, in Table 7.9, we find that the *dynamic* geo-based negative sampling approach can consistently improve the robustness of CRCF on Brightkite and Yelp dataset for the *Cold-Start Users* experiments. For example, CRCF$_{dgeo}$ obtains approximately 7% and 3% improvements in the Reward and Risk scores over DRCF on the Brightkite and Yelp datasets, respectively. Moreover, comparing between the T$_{risk}$ scores of CRCF and CRCF$_{dgeo}$ on the Brightkite dataset, at $\alpha = 5$, we observe that CRCF$_{dgeo}$ is less likely to exhibit a real risk of performing worse than the BPR baseline, while CRCF is not. Similarly, at $\alpha = 1$, on the Yelp dataset, we find that CRCF is likely to be under a real risk of performing worse than the BPR baseline, while CRCF$_{dgeo}$ is not. Next, the right hand plots in Figure 7.6 report that CRCF$_{dgeo}$ has consistently a larger number of positive changes in NDCG@10 over BPR across all positive bins in comparison with CRCF across the three datasets. These significant improvements in the T$_{risk}$ scores of CRCF$_{dgeo}$ compared to CRCF demonstrate that the *dynamic* geo-based negative sampling approach can significantly reduce the risk of CRCF framework in performing worse than BPR baselines. Furthermore, on analysing the right hand plots in Figure 7.7, we observe that CRCF$_{dgeo}$ consistently has larger number of positive changes in NDCG@10 over BPR across all positive bins in comparison with DRCF$_{dgeo}$ and CARA$_{dgeo}$ on the Brightkite and Yelp datasets. Overall, in response to research question RQ7.2(b), we find that the *dynamic* geo-based negative sampling approach can significantly reduce the risk of our proposed CRCF framework in performing worse than the BPR baseline for both the *Normal* and *Cold-Start* users experiments.

Next, we evaluate the usefulness of the *dynamic* geo-based negative sampling approach in enhancing the robustness of the CRCF framework using the T$_{risk}$ score. Similar to Figure 7.5, Figure 7.8 demonstrates the change in the T$_{risk}$ scores of the various approaches with the *dynamic* geo-based negative under *Normal* and *Cold-Start Users* experiments. With respect to the *Normal* experiments, as α increases, we observe that CRCF$_{dgeo}$ is significantly less risky than the BPR baseline until $\alpha = 7, \alpha = 12$ and $\alpha = 2$ for the Brightkite, Foursquare and Yelp datasets, respectively, while CRCF is not. These results suggest that the *dynamic* geo-based negative sampling approach can significantly improve the robustness of

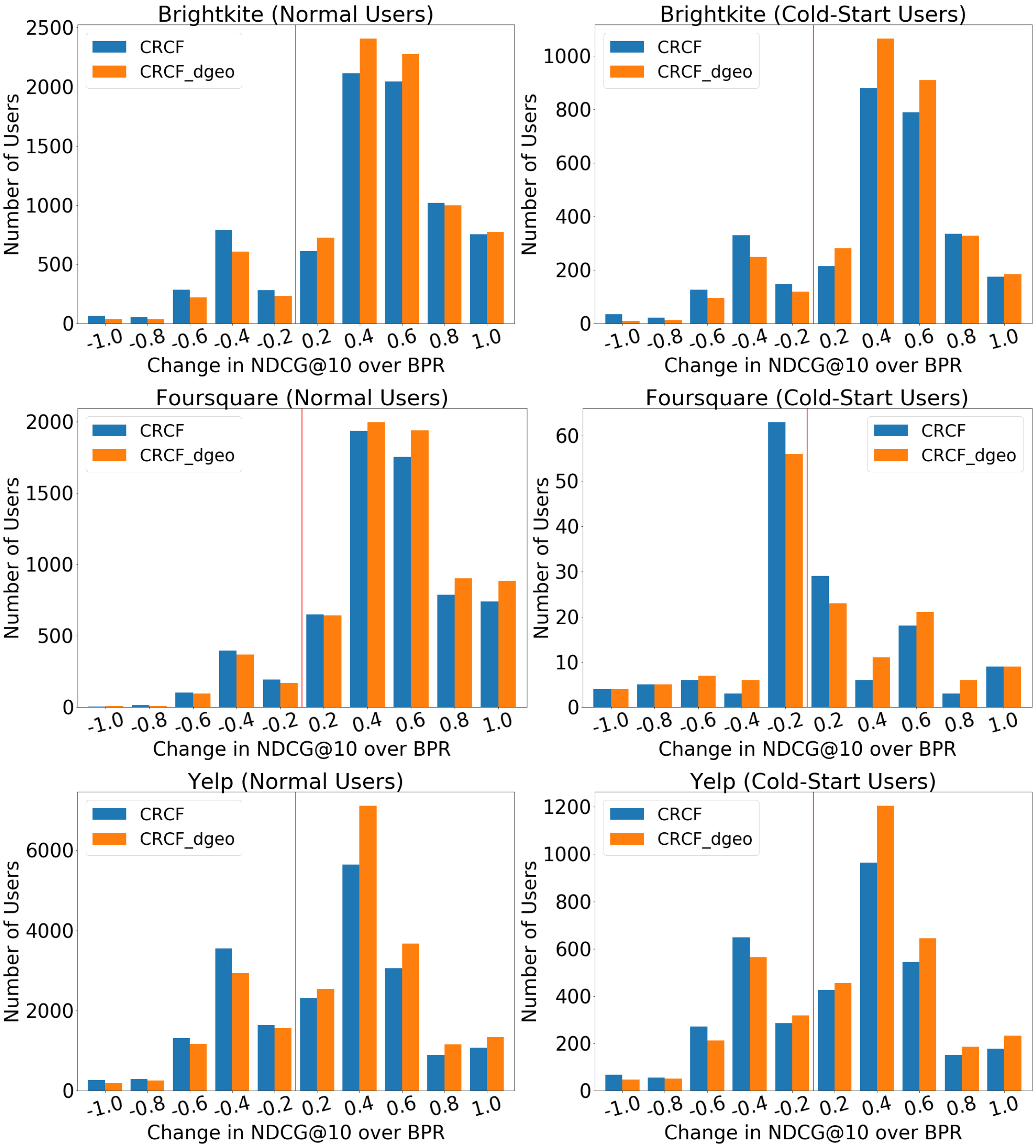


Figure 7.6: As per Figure 7.4, typical Wins-Losses histograms of the CRCF framework with or without the dynamic geo-based negative sampling approach (CRCF$_{dgeo}$ and CRCF) in comparison with the CF ranking baseline model, BPR, under the *Normal* and *Cold-Start Users* experiments.

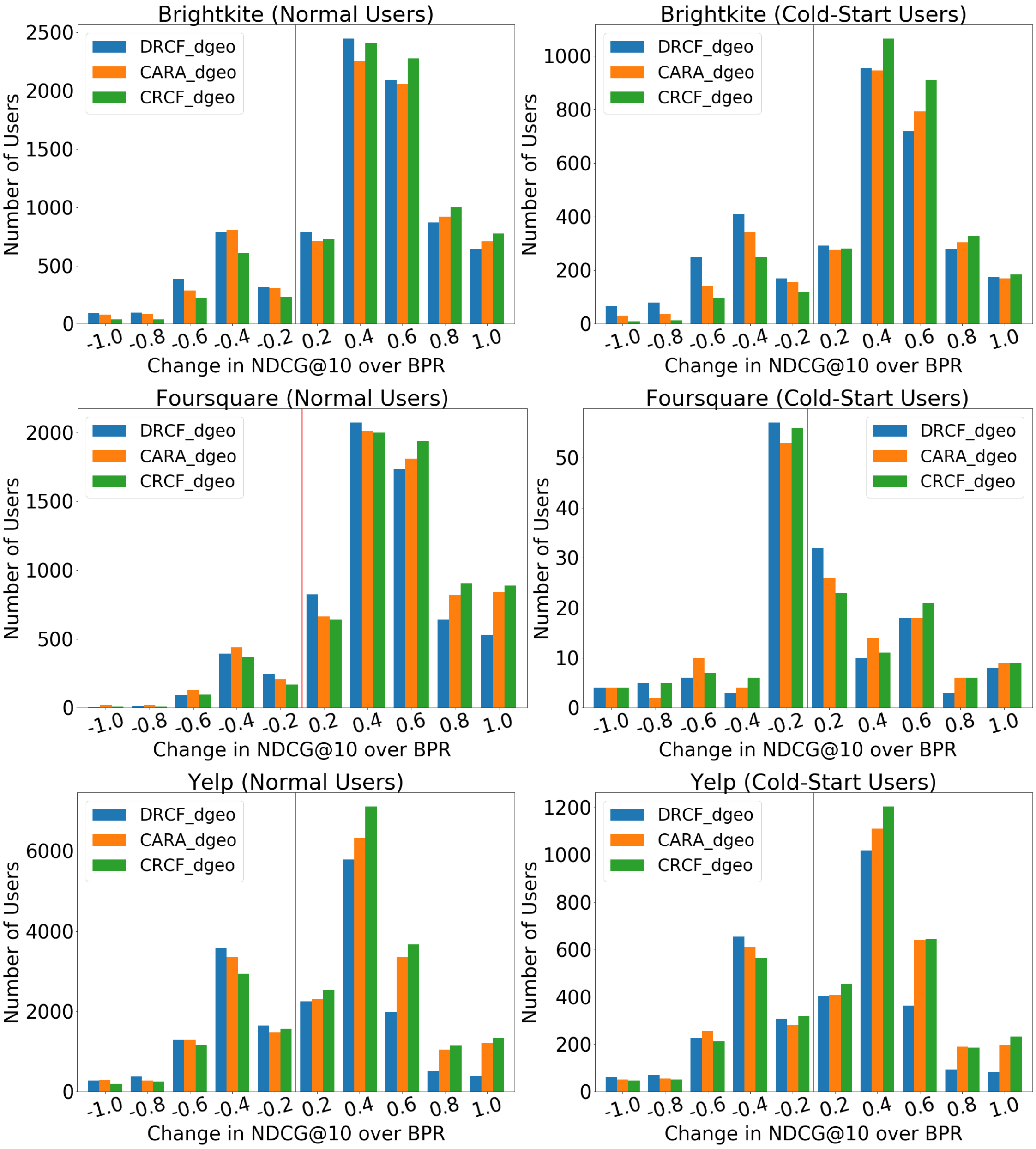


Figure 7.7: As per Figure 7.4, typical Wins-Losses histograms of target new models that incorporates the dynamic geo-based negative sampling approach (DRCF$_{dgeo}$, CARA$_{dgeo}$ and CRCF$_{dgeo}$) in comparison with the CF ranking baseline model, BPR, under the *Normal* and *Cold-Start Users* experiments.

our proposed CRCF framework (i.e. ensuring not to generate less effective recommendation than BPR). Overall, in response to research question RQ7.2(b), we observe further evidence that the *dynamic* geo-based negative sampling approach reduces the risk of proposed CRCF framework, for both *Normal* and *Cold-Start* users.

7.5 Conclusions

In this chapter, we proposed a novel Contextual Recurrent Collaborative Filtering framework (CRCF) for Context-Aware Venue Recommendation (CAVR). Our proposed CRCF framework is built on top of two deep neural network recommendation approaches, namely our proposed Deep Recurrent Collaborative Filtering (DRCF) framework, described in Chapter 5, and our proposed Contextual Attention Recurrent Architecture (CARA), described in Chapter 6. By combining DRCF and CARA together, CRCF aims to address **Limitation D1** of DRCF and **Limitation C1** of CARA. In particular, by exploiting both DRCF and CARA, CRCF can effectively capture the complex structure of the users' short-term (*dynamic*) and long-term (*static*) preferences by considering their preferred context (i.e. time of the day) as well as the contextual information associated with the sequence of the user's checkins.

Overall, our comprehensive experiments on three large-scale datasets from Brightkite, Foursquare and Yelp commercial LBSNs demonstrated the significant improvements of our proposed CRCF framework for CAVR in comparison with various existing state-of-the-art venue recommendation approaches in both the *Normal* and *Cold-Start* settings. In particular, in answering research question RQ7.1(a), our experimental results reported in Table 7.4 showed that CRCF significantly improved NDCG@10 by 5-20% over the DRCF framework and the CARA architecture across the three used datasets. These experimental results suggested that CRCF addresses both **Limitation D1** of the DRCF framework and **Limitation C1** of the CARA architecture. Moreover, in answering research question RQ7.2(a), our experimental results reported in Table 7.5 and Figure 7.3 demonstrated the usefulness of our proposed *dynamic* geo-based negative sampling approach, described in Section 5.3, (1) in enhancing the effectiveness of CRCF in generating high quality context-aware top-K venue recommendations and (2) in alleviating the cold-start problem. Moreover, our experimental results reported in Figures 7.2 showed that our proposed CRCF framework with the *dynamic* geo-based negative sampling approach converges faster than all baselines across the three used datasets.

Apart from the effectiveness of our proposed CRCF framework, we also investigated

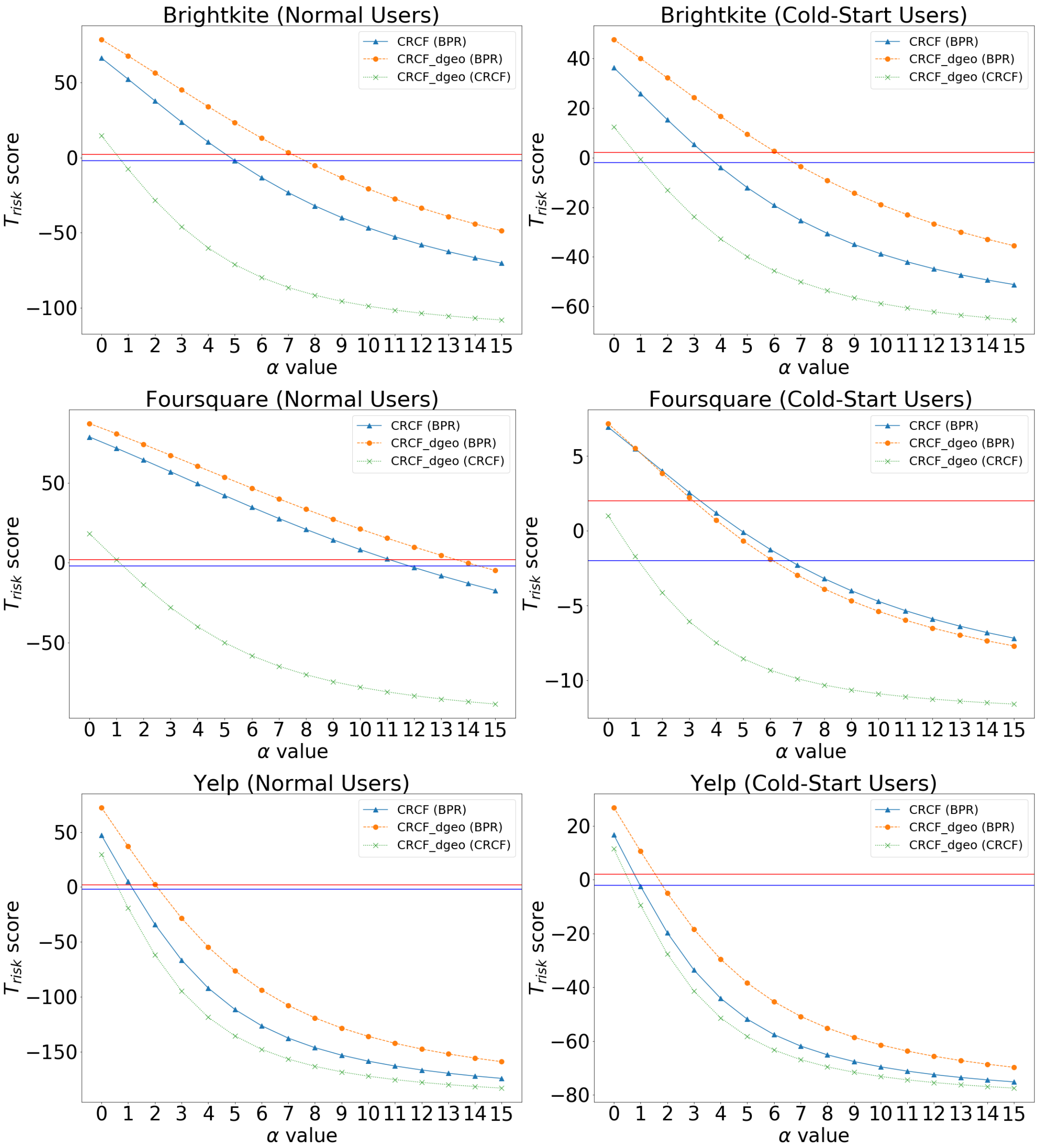


Figure 7.8: The changes in standardised T_{risk} scores for CRCF and CRCF$_{dgeo}$ with respect to the baseline, denoted inside the parentheses, over different α values under the *Normal* and *Cold-Start Users* experiments.

the robustness of CRCF in comparison with state-of-the-art factorisation approaches using risk-sensitive evaluation measures in Section 7.4.2. In answering research question RQ7.1(b), our experimental results reported in Tables 7.6 & 7.7 showed that our proposed CRCF framework is more robust (i.e. less likely to perform worse than the factorisation baseline (BPR)) than various factorisation approaches (e.g. DRCF and CARA) in both the *Normal* and *Cold-Start* settings. Our experimental results reported in Figures 7.4 demonstrated that it is highly likely that CRCF will not perform worse than the BPR baseline across the three used datasets, while both DRCF and CARA are more likely to perform worse than the BPR baseline on the Yelp dataset. Moreover, in respond to research question RQ7.2(b), we further investigated the robustness of CRCF when incorporates with the *dynamic* geo-based negative sampling approach. Our experimental results reported in Figures 7.6 and 7.7 demonstrated that the *dynamic* geo-based negative sampling approach can significantly improve the robustness of our proposed CRCF framework (i.e. ensuring not to generate less effective context-aware top-K venue recommendations than the BPR baseline) under the *Normal* and *Cold-Start* settings.

In our thesis statement (see Section 1.2), we hypothesised that, to generate high quality of context-aware top-K venue recommendations, a framework that consists of the following four functionalities/components is needed, namely (1) capturing the complex structure of the user-venue interactions in a collaborative filtering manner using an effective neural architecture to learn an arbitrary function from the user's implicit feedback, (2) modelling the user's long- (*static*) and short-term (*dynamic*) preferences from the sequential order of user's checkins and the contextual information associated with the successive checkins, (3) generating accurate top-K venue suggestions based on the user's *static* and *dynamic* preferences using a pairwise ranking function and (4) sampling potential negative instances that take into account the additional information such as geographical information of venues, users' social relationships and the sequential order of users' checkins. Based upon our effectiveness experiments in this chapter, we conclude that our proposed CRCF framework, which consists of the aforementioned four functionalities/components, can improve the quality of context-aware top-K venue recommendations. Furthermore, based upon our robustness experiments, we conclude that the CRCF framework is robust and less likely to generate less effective context-aware top-K venue recommendations than the BPR baseline. In the next chapter, we close this thesis by summarising the conclusions and outcomes from each of the individual chapters, in addition to providing possible new research directions uncovered by this work.

Table 7.4: Performance in terms of HR@10 and NDCG@10 between various approaches. The best performing approach is highlighted in bold; $-$ and $*$ denote a significant difference compared to the best performing result, according to the paired t-test for $p < 0.05$ and $p < 0.01$, respectively.

Normal Users Experiments						
	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
MostPop	0.1462*	0.1010*	0.2009*	0.1167*	0.0739*	0.0334*
MostVisit	0.4032*	0.3473*	0.4733*	0.4290*	0.1083*	0.0528*
RecentVisit	0.4809*	0.4370*	0.4584*	0.4037*	0.1096*	0.0542*
MF	0.6206*	0.3470*	0.6656*	0.3818*	0.3539*	0.1734*
RNN	0.6368*	0.3824*	0.8040*	0.5459*	0.3814*	0.1891*
BPR	0.6890*	0.4333*	0.7550*	0.4834*	0.4963*	0.2676*
DREAM	0.7041*	0.4839*	0.8147*	0.6081*	0.4349*	0.2235*
STELLAR	0.7267*	0.5635*	0.8751*	0.6984*	0.5356*	0.2969*
NeuMF	0.7073*	0.5358*	0.8361*	0.5842*	0.4934*	0.2729*
DRCF	0.7419*	0.6048*	0.8952*	0.7223*	0.5162*	0.2963*
CARA	0.7385*	0.6040*	0.8851*	0.7154*	0.5587*	0.3272*
CRCF	**0.7528**	**0.6319**	**0.8981**	**0.7442**	**0.5861**	**0.3479**
Cold-Start Users Experiments						
	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NCDG	HR	NDCG
MostPop	0.1155*	0.0778*	0.0584*	0.0286*	0.0714*	0.0316*
MostVisit	0.4285*	0.3789*	0.3506*	0.3175*	0.1044*	0.0489*
RecentVisit	0.4995*	0.4585*	0.3831*	0.3446*	0.1052*	0.0497*
MF	0.6768*	0.3913*	0.6623*	0.3650*	0.3748*	0.1868*
BPR	0.7519*	0.4907*	0.7792-	0.4961*	0.5273*	0.2946*
RNN	0.6486*	0.3694*	0.5909*	0.4041*	0.3856*	0.1901*
DREAM	0.7452*	0.4969*	0.7987-	0.5379*	0.4523*	0.2239*
STELLAR	0.7406*	0.5580*	0.8052-	0.6007*	0.5537*	0.3147*
NeuMF	0.7160*	0.5894*	0.7922-	0.6227*	0.5102*	0.2734*
DRCF	0.7526*	0.5980*	0.8377	0.6645	0.5330*	0.3136*
CARA	0.7648*	0.6220*	**0.8636**	0.6505-	0.5748*	0.3493*
CRCF	**0.7782**	**0.6582**	0.8571	**0.6967**	**0.5913**	**0.3622**

Table 7.5: As per Table 7.4. Performances in terms of HR@10 and NDCG@10 between various approaches that apply our proposed *dynamic* geo-based negative sampling approach (see Section 5.3), denoted as $dgeo$

Normal Users Experiments						
	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
MostPop	0.1462*	0.1010*	0.2009*	0.1167*	0.0739*	0.0334*
MostVisit	0.4032*	0.3473*	0.4733*	0.4290*	0.1083*	0.0528*
RecentVisit	0.4809*	0.4370*	0.4584*	0.4037*	0.1096*	0.0542*
GeoBPR	0.7339*	0.4672*	0.8216*	0.5395-	0.5570*	0.3032*
DRCF	0.7419*	0.6048*	0.8952*	0.7223*	0.5162*	0.2963*
CARA	0.7385*	0.6040*	0.8851*	0.7154*	0.5587*	0.3272*
CRCF	0.7528	0.6319	0.8981	0.7442	0.5861	0.3479
DRCF$_{dgeo}$	0.7852*	0.6210*	0.9095*	0.7214*	0.5618*	0.3064*
CARA$_{dgeo}$	0.7717*	0.6266*	0.9129*	0.7567*	0.6107*	0.3665*
CRCF$_{dgeo}$	**0.8029**	**0.6606**	**0.9260**	**0.7788**	**0.6548**	**0.3927**
Cold-Start Users Experiments						
	Brightkite		Foursquare		Yelp	
Model	HR	NDCG	HR	NDCG	HR	NDCG
MostPop	0.1155*	0.0778*	0.0584*	0.0286*	0.0714*	0.0316*
MostVisit	0.4285*	0.3789*	0.3506*	0.3175*	0.1044*	0.0489*
RecentVisit	0.4995*	0.4585*	0.3831*	0.3446*	0.1052*	0.0497*
GeoBPR	0.8093*	0.5262*	0.8312-	0.5486*	0.5802*	0.3202*
DRCF	0.7526*	0.5980*	0.8377*	0.6645*	0.5330*	0.3136*
CARA	0.7648*	0.6220*	0.8636*	0.6505*	0.5748*	0.3493*
CRCF	0.7782*	0.6582*	0.8571*	0.6967*	0.5913*	0.3622*
DRCF$_{dgeo}$	0.8094*	0.6199*	**0.8896**	0.7074	0.5877*	0.3318*
CARA$_{dgeo}$	0.8153*	0.6556*	0.8766	**0.7225**	0.6332*	0.3893*
CRCF$_{dgeo}$	**0.8557**	**0.6995**	0.8701	0.7152	**0.6612**	**0.4053**

Table 7.6: The robustness of various approaches in comparison with the BPR baseline in terms of NDCG@10 on three datasets for *Normal* users. T$_{risk}$ scores greater than $+2$ or less than -2 indicate that a two-tailed paired t-test gives significance with $p < 0.05$. T$_{risk}$ scores greater than $+2$ are indicated with $*$. The best score w.r.t. each risk-sensitive measure is highlighted in bold.

Dataset	Measure	DRCF	CARA	CRCF
Brightkite	Risk	0.052	0.049	**0.041**
	Reward	0.224	0.220	**0.240**
	Wins/Losses	6297/1805	6159/1687	**6538/1483**
	Wins%/Losses%	43.80/12.55	42.84/11.73	**45.48/10.31**
	U$_{risk}$ $\alpha = 1$	0.119	0.122	0.157
	T$_{risk}$ $\alpha = 1$	38.763*	40.390*	52.161*
	U$_{risk}$ $\alpha = 5$	-0.090	-0.073	-0.007
	T$_{risk}$ $\alpha = 5$	-24.066	-20.195	-2.051
Foursquare	Risk	**0.019**	0.029	0.022
	Reward	0.258	0.261	**0.283**
	Wins/Losses	5723/**644**	5450/910	**5876**/711
	Wins%/Losses%	53.14/**5.98**	50.61/8.45	**54.56**/6.60
	U$_{risk}$ $\alpha = 1$	0.220	0.202	0.238
	T$_{risk}$ $\alpha = 1$	71.063*	60.024*	71.698*
	U$_{risk}$ $\alpha = 5$	0.142	0.084	0.147
	T$_{risk}$ $\alpha = 5$	44.010*	22.915*	42.049*
Yelp	Risk	0.071	0.075	**0.070**
	Reward	0.097	0.133	**0.148**
	Wins/Losses	9232/7311	11820/7518	**12980/7064**
	Wins%/Losses%	23.70/18.77	30.35/19.30	**33.32/18.13**
	U$_{risk}$ $\alpha = 1$	-0.045	-0.017	0.009
	T$_{risk}$ $\alpha = 1$	-30.236	-10.322	5.025*
	U$_{risk}$ $\alpha = 5$	-0.330	-0.320	-0.272
	T$_{risk}$ $\alpha = 5$	-142.903	-126.541	-111.453

Table 7.7: The robustness of various approaches in comparison with the BPR baseline in terms of NDCG@10 on three datasets for the *Cold-Start* users. T$_{risk}$ scores greater than $+2$ or less than -2 indicate that a two-tailed paired t-test gives significance with $p < 0.05$. T$_{risk}$ scores greater than $+2$ are indicated with $*$. The best score w.r.t. each risk-sensitive measure is highlighted in bold.

Dataset	Measure	DRCF	CARA	CRCF
Brightkite	Risk	0.082	0.060	**0.047**
	Reward	0.190	0.192	**0.215**
	Wins/Losses	2183/1043	2221/803	**2394/661**
	Wins%/Losses%	39.13/18.69	39.81/14.39	**42.91/11.85**
	U$_{risk}$ $\alpha = 1$	0.025	0.071	0.121
	T$_{risk}$ $\alpha = 1$	4.871*	15.045*	25.832*
	U$_{risk}$ $\alpha = 5$	-0.304	-0.170	-0.067
	T$_{risk}$ $\alpha = 5$	-40.821	-27.581	-12.069
Foursquare	Risk	0.056	0.067	**0.041**
	Reward	0.224	0.221	**0.241**
	Wins/Losses	**73/20**	57/86	65/81
	Wins%/Losses%	**47.40/12.98**	37.01/55.84	42.20/52.59
	U$_{risk}$ $\alpha = 1$	0.113	0.087	0.160
	T$_{risk}$ $\alpha = 1$	3.785*	2.813*	5.498*
	U$_{risk}$ $\alpha = 5$	-0.109	-0.181	-0.003
	T$_{risk}$ $\alpha = 5$	-2.955	-4.438	-0.092
Yelp	Risk	0.079	0.079	**0.078**
	Reward	0.098	0.134	**0.145**
	Wins/Losses	1593/1421	2119/1374	**2266/1328**
	Wins%/Losses%	23.07/20.58	30.69/19.90	**32.82/19.23**
	U$_{risk}$ $\alpha = 1$	-0.059	-0.024	-0.010
	T$_{risk}$ $\alpha = 1$	-16.220	-5.982	-2.411
	U$_{risk}$ $\alpha = 5$	-0.374	-0.341	-0.321
	T$_{risk}$ $\alpha = 5$	-63.249	-55.016	-51.850

Table 7.8: The robustness of various approaches that incorporate the *dynamic* geo-based negative sampling in comparison with the BPR baseline in terms of NDCG@10 on three datasets for *Normal* users. T$_{risk}$ scores greater than $+2$ or less than -2 indicate that a two-tailed paired t-test gives significance with $p < 0.05$. T$_{risk}$ scores greater than $+2$ are indicated with $*$. The best score w.r.t. each risk-sensitive measure is highlighted in bold.

Dataset	Measure	DRCF	CARA	CRCF	DRCF$_{dgeo}$	CARA$_{dgeo}$	CRCF$_{dgeo}$
Brightkite	Risk	0.052	0.049	0.041	0.049	0.044	**0.031**
	Reward	0.224	0.220	0.240	0.237	0.238	**0.258**
	Wins/Losses	6297/1805	6159/1687	6538/1483	6836/1677	6660/1568	**7179/1139**
	Wins%/Losses%	43.80/12.55	42.84/11.73	45.48/10.31	47.55/11.66	46.33/10.90	**49.94/7.92**
	U$_{risk}$ $\alpha = 1$	0.119	0.122	0.157	0.138	0.149	0.196
	T$_{risk}$ $\alpha = 1$	38.763*	40.390*	52.161*	45.453*	49.212*	67.637*
	U$_{risk}$ $\alpha = 5$	-0.090	-0.073	-0.007	-0.059	-0.029	0.073
	T$_{risk}$ $\alpha = 5$	-24.066	-20.195	-2.051	-16.113	-8.170	23.171*
Foursquare	Risk	**0.019**	0.029	0.022	0.022	0.027	0.021
	Reward	0.258	0.261	0.283	0.260	0.301	**0.316**
	Wins/Losses	5723/**644**	5450/910	5876/711	5798/749	6153/821	**6371**/652
	Wins%/Losses%	43.80/12.55	42.84/11.73	45.48/10.31	53.84/6.95	57.14/7.62	**59.16/6.05**
	U$_{risk}$ $\alpha = 1$	0.220	0.202	0.23	0.215	0.245	0.274
	T$_{risk}$ $\alpha = 1$	71.063*	60.024*	71.698*	68.998*	70.814*	80.755*
	U$_{risk}$ $\alpha = 5$	0.142	0.084	0.147	0.125	0.134	0.189
	T$_{risk}$ $\alpha = 5$	44.010*	22.915*	42.049*	37.960*	36.110*	53.524*
Yelp	Risk	0.071	0.075	0.070	0.072	0.067	**0.059**
	Reward	0.097	0.133	0.148	0.109	0.165	**0.183**
	Wins/Losses	9232/7311	11820/7518	12980/7064	10927/7184	14270/6721	**15807/6117**
	Wins%/Losses%	23.70/18.77	30.35/19.30	33.32/18.13	28.05/18.44	36.64/17.25	**40.58/15.70**
	U$_{risk}$ $\alpha = 1$	-0.045	-0.017	0.009	-0.035	0.029	0.064
	T$_{risk}$ $\alpha = 1$	-30.236	-10.322	5.025*	-22.989	16.783*	36.938*
	U$_{risk}$ $\alpha = 5$	-0.330	-0.320	-0.272	-0.323	-0.243	-0.174
	T$_{risk}$ $\alpha = 5$	-142.903	-126.541	-111.453	-137.942	-99.610	-76.218

Table 7.9: The robustness of various approaches that incorporate the *dynamic* geo-based negative sampling in comparison with the BPR baseline in terms of NDCG@10 on three datasets for the *Cold-Start* users. T$_{risk}$ scores greater than $+2$ or less than -2 indicate that a two-tailed paired t-test gives significance with $p < 0.05$. T$_{risk}$ scores greater than $+2$ are indicated with $*$. The best score w.r.t. each risk-sensitive measure is highlighted in bold.

Dataset	Measure	DRCF	CARA	CRCF	DRCF$_{dgeo}$	CARA$_{dgeo}$	CRCF$_{dgeo}$
Brightkite	Risk	0.082	0.060	0.047	0.078	0.051	**0.033**
	Reward	0.190	0.192	0.215	0.207	0.216	**0.241**
	Wins/Losses	2183/1043	2221/803	2394/661	2418/974	2489/706	**2767/486**
	Wins%/Losses%	39.13/18.69	39.81/14.39	42.91/11.85	43.34/11.66	44.62/10.90	**49.60/7.92**
	U$_{risk}$ $\alpha = 1$	0.025	0.071	0.121	0.051	0.114	0.176
	T$_{risk}$ $\alpha = 1$	4.871*	15.045*	25.832*	9.871*	24.273*	39.980*
	U$_{risk}$ $\alpha = 5$	-0.304	-0.170	-0.067	-0.262	-0.090	0.046
	T$_{risk}$ $\alpha = 5$	-40.821	-27.581	-12.069	-35.896	-15.663	9.396*
Foursquare	Risk	0.056	0.067	**0.041**	0.043	0.049	0.049
	Reward	0.224	0.221	0.241	0.254	**0.275**	0.268
	Wins/Losses	**73/20**	57/86	65/81	**71**/75	**73**/73	70/78
	Wins%/Losses%	**47.40/12.98**	37.01/55.84	42.20/52.59	46.10/48.70	**47.40**/47.40	45.45/50.64
	U$_{risk}$ $\alpha = 1$	0.113	0.087	0.160	0.168	0.178	0.170
	T$_{risk}$ $\alpha = 1$	3.785*	2.813*	5.498*	5.825*	5.796*	5.536*
	U$_{risk}$ $\alpha = 5$	-0.109	-0.181	-0.003	-0.004	-0.018	-0.024
	T$_{risk}$ $\alpha = 5$	-2.955	-4.438	-0.092	-0.115	-0.491	-0.674
Yelp	Risk	0.079	0.079	0.078	0.075	0.072	**0.066**
	Reward	0.098	0.134	0.145	0.113	0.166	**0.177**
	Wins/Losses	1593/1421	2119/1374	2266/1328	1962/1320	2548/1257	**2724/1195**
	Wins%/Losses%	23.07/20.58	30.69/19.90	32.82/19.23	28.42/19.12	36.91/18.20	**39.46/17.31**
	U$_{risk}$ $\alpha = 1$	-0.059	-0.024	-0.010	-0.038	0.023	0.045
	T$_{risk}$ $\alpha = 1$	-16.220	-5.982	-2.411	-10.301	5.468*	10.617*
	U$_{risk}$ $\alpha = 5$	-0.374	-0.341	-0.321	-0.339	-0.264	-0.220
	T$_{risk}$ $\alpha = 5$	-63.249	-55.016	-51.850	-58.746	-44.224	-38.381

Chapter 8

Conclusions

8.1 Contributions and Conclusions

In this thesis, we addressed the challenges of Context-Aware Venue Recommendation (CAVR), namely, (1) modelling the users' preferences and the characteristics of venues from the users' explicit feedback (e.g. the users' ratings and comments), (2) modelling the users' long- (*static*) and short-term (*dynamic*) preferences from the sequential order of the user's implicit feedback (e.g. checkins) and the contextual information associated with the successive feedback, (3) generating accurate top-K venue suggestions based on the user's *static* and *dynamic* preferences using a pairwise ranking function and (4) sampling potential negative instances that take into account additional information such as the geographical information of venues, the users' social relationships and the sequential order of users' checkins. In particular, in Chapter 4, we proposed the Social and Textual Regularisation technique (STReg) and the textual Matrix Factorisation-based approach (MFw2v) that leverage the textual content of users' comments to enhance the effectiveness of the traditional MF approach in modelling the users' preferences and the characteristics of venues. In addition, we proposed the Personalised Ranking Framework with Multiple sampling Criteria framework (PRFMC) that could incorporate multiple sources of additional information to effectively sample negative instances. In Chapter 5, we proposed the Deep Recurrent Collaborative Filtering framework (DRCF) that could effectively model the users' *static* and *dynamic* preferences from the users' sequence of checkins. We empirically showed that our proposed DRCF framework could generate more effective top-K venue recommendations than existing Deep Neural Network-based approaches. To incorporate the contextual information associated with the users' sequence of checkins, in Chapter 6, we proposed the Contextual Attention Recurrent Architecture (CARA) for CAVR, which could effectively incorporate different types of contextual information to model the users' contextual *dynamic* preferences. Our experimental results in Chapter 6 showed that our proposed CARA architecture could generate more effective context-aware venue recommendations than various previous state-of-the-art recurrent architectures. Finally, in Chapter 7, we integrated our proposed CARA architecture into our proposed DRCF framework to further enhance the quality of context-aware venue recommendations.

In the remainder of this chapter, we first summarise the contributions of this thesis, which are in Section 8.1.1 followed by the main achievements and conclusions drawn from this thesis, which are presented in Section 8.1.2. We discuss some future research directions for context-aware venue recommendations in Section 8.2. Finally, we present our closing remarks in Section 8.3.

8.1.1 Contributions

The main contributions of this thesis are as follows:

- In Chapter 4, we proposed the novel Social and Textual Regularisation (STReg) technique and the textual MF-based approach (MFw2v) that exploit word embeddings to model the semantic properties of the textual content of comments associated with the users' ratings to enhance the effectiveness of the traditional MF model for the user-venue rating prediction. To evaluate the effectiveness of our proposed STReg technique and our MFw2v approach, we conducted experiments on a large-scale rating dataset from the Yelp LBSN, which consists of over 2.2 million ratings from 500k users. We compared the effectiveness of our proposed STReg technique and MFw2v approach with the existing MF-based approaches (e.g. CMF and JMF). Moreover, in this chapter, we proposed a novel Personalised Ranking Framework with Multiple sampling Criteria framework (PRFMC), which can leverage multiple types of additional information to enhance the effectiveness of the traditional BPR model for top-K venue recommendations. To evaluate the effectiveness of PRFMC for top-K venue recommendations, we conducted several experiments using publicly available large-scale LBSN datasets. We measured the quality of the top-K venue recommendations in terms of ranking-based metrics, which were previously described in Section 2.1.2.2.
- In Chapter 5, we proposed the novel Deep Recurrent Collaborative Filtering framework (DRCF), an extension of the NeuMF framework (described in Section 3.4), which exploits Recurrent Neural Networks (RNN) to capture the users' *dynamic* preferences from their sequences of checkins. The DRCF framework consists of three

components: namely (i) a Generalised Recurrent Matrix Factorisation (GRMF) model, (ii) a Multi-Layer Recurrent Perceptron (MLRP) model and (iii) a Recurrent Matrix Factorisation (RMF) model. Within the DRCF framework, we proposed novel *dynamic* and *static* geo-based negative sampling approaches that take the sequential properties of checkins and geographical location of venues into account to enhance the effectiveness of the DRCF framework, as well as alleviate the cold-start user problem. We conducted several experiments to evaluate the effectiveness of the DRCF framework using publicly available large-scale checkin and rating datasets from the Brightkite, Foursquare and Yelp LBSNs. We compared the effectiveness of our proposed DRCF framework and its components with various existing baselines, which could be categorised into traditional MF-based approaches, RNN-based approaches and Deep Neural Network-based approaches (see Section 5.7.2).

- In Chapter 6, we proposed the novel Contextual Attention Recurrent Architecture (CARA) for context-aware venue recommendations that could effectively incorporate different types of contextual information associated with the users' sequence of checkins. In particular, our proposed CARA architecture consisted of two types of gating mechanisms: namely a Contextual Attention Gate (CAG) as well as Temporal and Spatial Gates (TSG). The CAG gate aims to effectively capture the users' contextual *dynamic* (short-term) preferences by taking into account the *ordinary* context associated with the users' checkins, while the TSG gates aim to capture the correlation between the users' previous checkin and the current checkin from the *transition* context associated with two successive checkins. In Section 6.6, we evaluated the effectiveness of our proposed CARA architecture in comparison with state-of-the-art GRU architectures and RNN-based factorisation approaches using three large-scale checkin and rating datasets.

- In Chapter 7, we proposed a novel Contextual Recurrent Collaborative Filtering framework (CRCF) that combines the DRCF framework proposed in Chapter 5 and the CARA architecture proposed in Chapter 6. The CRCF framework leverages the sequence of the users' checkins, the users' preferred context and the contextual information associated with the sequence of the users' checkins to effectively and comprehensively capture the users' contextual long-term (*static*) and short-term (*dynamic*) preferences for context-aware venue recommendations. We empirically evaluated the effectiveness of the CRCF framework in comparison with the existing factorisation approaches as well as both our proposed DRCF framework and our proposed CARA architecture. Moreover, we investigated the robustness of the CRCF framework by leveraging risk analysis techniques proposed by Wang et al. (2012); Dinçer et al. (2014).

8.1.2 Conclusions

In this section, we summarise the main conclusions and achievements of this thesis. Then, we validate our thesis statement proposed in Section 1.2 based on our main conclusions and achievements.

- **Effectiveness of Word Embeddings for User-Venue Rating Prediction:** In Chapter 4, we proposed the Social and Textual Regularisation (STReg) technique and the textual MF-based approach (MFw2v) that exploit word embeddings to model the semantic properties of the textual content of comments associated with the users' rating to enhance the effectiveness of the traditional MF model for the user-venue rating prediction. In Section 4.2.4, we empirically evaluated the effectiveness of STReg and MFw2v on the Yelp user-venue rating dataset. We demonstrated the usefulness of our proposed STReg technique and our MFw2v approach in improving the prediction accuracy of the traditional MF model (see Table 4.3). By exploiting word embeddings to extract the semantic properties of the users' comments, both the STReg technique and the MFw2v approach are more effective than Bag-of-Words MF-based baselines. In particular, the results in Table 4.2 show that both STReg and MFw2v can outperform the Bag-of-Words MF-based baselines by 6-16% in terms of MAE and RMSE on the Yelp dataset. Hence, we concluded that our proposed STReg technique and our MFw2v approach were effective in predicting the users' ratings on venues.

- **Effectiveness of Multiple Sampling Criteria for Top-K Venue Recommendations:** In Chapter 4, we proposed the Personalised Ranking Framework with Multiple sampling Criteria (PRFMC) that incorporates multiple types of additional information to further improve the quality of top-K venue recommendations of the BPR model. Experimental results in Section 4.3.6 showed that our proposed PRFMC framework, which leverages both the users' geographical movements and the users' social influences, could effectively sample negative instances and improve the quality of top-K venue recommendations compared to various state-of-the-art BPR-based approaches (see Section 4.3.5). In particular, we found that our PRFMC framework provided significant benefits across the Yelp, Brightkite and Gowalla datasets in terms of MAP, NDCG and MRR, compared to various existing state-of-the-art single criterion negative sampling approaches, multiple criteria negative sampling approaches as well as probabilistic models (see Table 4.6 and Table 4.7).

- **Usefulness of a Recurrent Neural Network for Sequence-Aware Venue Recommendation:** In Chapter 5, we proposed a novel Deep Recurrent Collaborative Filtering

(DRCF) framework, which exploits Recurrent Neural Networks (RNN) to effectively capture the users' *dynamic* preferences from their sequences of checkins. In Section 5.8.1, we showed that the sequential order of users' checkins plays an important role in enhancing the quality of venue recommendation. The experimental results on three large-scale checkin datasets demonstrated the effectiveness of our DRCF framework that exploits RNN to capture the users' *dynamic* (short-term) preferences for venue recommendation compared to various RNN-based approaches (see Table 5.4). In Section 5.8.2, we showed that the neural architectures such as the Multi-Layer Perceptron were more effective in capturing the complex structure of user-venue interactions than the traditional MF-based approaches (see Table 5.5). Moreover, within the DRCF framework, we proposed the novel *dynamic* and geo-based negative sampling approach that takes the users' sequential properties of checkins and the geographical information of venues into account during the negative sampling process. Through our comprehensive analysis in Section 5.8.3 , we showed that our proposed *dynamic* negative sampling approach can significantly enhance the effectiveness of the DRCF framework in terms of HR by 5.8% and 1.6% on the Brightkite and Foursquare datasets, respectively (see Table 5.6). Furthermore, in Section 5.8.4, we showed that our proposed *dynamic* negative sampling approach can alleviate the cold-start user problem by significantly improving the effectiveness of the DRCF framework under the cold-start setting by approximately 3.5% and 6.5% in terms of NDCG on the Brightkite and Foursquare datasets, respectively (see Table 5.7).

- **Effectiveness of the Recurrent Architecture for Context-Aware Venue Recommendation:** In Chapter 6, we proposed the Contextual Attention Recurrent Architecture (CARA) that could effectively capture the users' contextual *dynamic* (short-term) preferences from the users' sequence of checkins. In Section 6.7, we showed that the contextual information associated with the sequences of users' checkins such as the time of the day as well as the time intervals and the distances between two successive checkins play an important role in enhancing the quality of context-aware venue recommendation. In particular, we demonstrated that our CARA architecture, which incorporates the contextual information significantly outperformed various factorisation-based approach and state-of-the-art recurrent architectures by 5-18% and 6-27% in terms of HR and NDCG across the Brightkite, Foursquare and Yelp datasets (see Tables 6.4-6.7). Furthermore, in Section 6.8, we showed that, the users' *dynamic* preferences in LBSNs were determined by the *transition* context extracted from the time intervals and the geographical distances between two successive checkins (see Figure 6.3).

- **Effectiveness of the Contextual Recurrent Collaborative Filtering Framework for Context-Aware Venue Recommendation:** In Chapter 7, we showed that our proposed Contextual Recurrent Collaborative Filtering (CRCF) framework, which combines our proposed DRCF framework (see Chapter 5) and our proposed CARA architecture (see Chapter 6), can effectively capture the users' *static* (long-term) and *dynamic* (short-term) preferences by considering their preferred context (i.e. time of the day) as well as the contextual information associated with the sequence of the user's checkins. The experimental results on the three used large-scale LBSNS datasets reported in Section 7.4.1 showed that CRCF significantly improved the ranking metrics over the DRCF framework and the CARA architecture across the three used datasets (see Tables 7.4-7.5 and Figures 7.2-7.3). Furthermore, in Section 7.4.2, we showed that our proposed CRCF framework is robust, as it is less likely to generate low quality context-aware venue recommendations than traditional BPR and other baselines (see Tables 7.6-7.7 and Figures 7.4-7.5)

Next, we validate our thesis statement, proposed in Section 1.2, based on our empirical studies in Chapters 4-7. In summary, the key statement of this thesis is that the quality of context-aware venue recommendation, could be effectively enhanced by leveraging the additional information such as the users' social relationships, the textual content of comments, the geographical information of venues and the sequential properties of the users' checkin feedback as well as the contextual information associated with the sequences of users' checkin feedback.

- We claimed that leveraging the users' social information and the textual content of comments could effectively enhance the user-venue rating prediction accuracy of the traditional MF approach. We argue that we have validated this claim in Chapter 4 where we showed that regularising the traditional MF approach based on the semantic similarity between the users and their friends extracted from their comments can improve the prediction accuracy of the traditional MF approach (see Table 4.3). Furthermore, we also showed that factorising the semantic properties of the users' and venues' comments can improve the accuracy of user-venue rating prediction by 5-16% (see Table 4.3). Overall, these empirical studies showed that both the users' social information and the textual content of users' comments play an important role in enhancing the prediction accuracy of traditional MF approach.

- We postulated that the quality of personalised top-K venue recommendation could be improved by negative sampling processes that take into account the users' social information and the geographical information of venues. Our experiments in Chapter 4 validated this claim by showing that our proposed Personalised Ranking Framework with Multiple sampling Criteria (PRFMC) that incorporates the users' social relationship and the geographical location of venues could significantly improve the quality of top-K venue recommendation (according to paired t-test with $p < 0.01$) in comparison with various state-of-the-art BPR-based approaches (see Table 4.6 and Table 4.7).

- We claimed that leveraging the sequential order of the users' checkin feedback to model the users' long- (*static*) and short-term (*dynamic*) preferences could improve the quality of personalised top-K venue recommendation. To validate this claim, in Chapter 5, we showed that exploiting deep neural network and recurrent model to incorporate the sequential order of users' checkins into the traditional Collaborative Filtering-based approach led to signifiant improvement in top-K venue recommendations, according to paired t-test with $p < 0.01$ (see Table 5.4 and Figure 5.2). In addition, we further validated this claim by showing that negative sampling approaches that take the sequential properties of users' checkins into account could further enhance the quality of top-K venue recommendation (see Tables 5.6 & 5.7)

- We claimed that leveraging the contextual information associated with the sequences of users' checkins to model the users' short-term (*dynamic*) preferences could enhance the quality of context-aware venue recommendation. We argue that we have validated this claim in Chapter 7 where we showed that the recurrent architectures that incorporate the contextual information associated with the sequences of users' checkins (e.g. the time interval and distance between two successive checkins) to model the users' *dynamic* preferences can significantly improve the quality of context-aware venue recommendation (see Tables 6.4-6.6 and Figure 6.3).

- We postulated that leveraging the geographical information of venues, the sequential order of the users' checkin feedback and the contextual information associated with the sequences of users' checkins to comprehensively model the users' long- (*static*) and short-term (*dynamic*) preferences could improve the quality of context-aware venue recommendation. Our experiments in Chapter 7 showed that our proposed Contextual Recurrent Collaborative Filtering framework (CRCF) that combines our proposed DRCF framework (Chapter 5) and our CARA architecture (Chapter 6) to comprehensively leverage the aforementioned additional information leads to a significant improvement in context-aware venue recommendation (see Tables 7.4-7.7 and Figures 7.2-7.5). Hence, we concluded that the geographical information of venues, the sequential order of the users' checkin feedback and the contextual information associated with the sequences of users' checkins are useful in enhancing the quality of context-aware venue recommendation.

Overall, we argue that we have validated all claims of our thesis statement using different three large-scale datasets. In particular, we showed that our proposed CRCF framework, which exploits deep neural networks, recurrent architectures as well as negative sampling processes that incorporate the geographical information of venues, the sequential properties of users' checkin feedback and the contextual information associated with the sequences of users' checkin feedback significantly improves the quality of context-aware venue recommendation.

8.2 Directions for Future Work

In this section, we discuss possible directions for future research related to context-aware venue recommendation. In particular, we discuss future research directions that have become apparent as a direct result of the work that we have presented in this thesis.

Modelling the Users' Preferences and Characteristics of Venues from Textual Information: In Chapter 4, we showed that the word embeddings are effective to represent the semantic properties of the users' textual comments, which could be exploited to effectively model the users' preferences and the characteristics of venues. In particular, to represent a user's preference from his/her comments, our proposed Social and Textual Regularisation technique (STReg) and our textual MF-based approach (MFw2v) sum the word embedding representations of each term occurred in the user's comment. However, such approach ignores the position of terms, which is not effective to capture the semantic properties of the comments. Various deep learning techniques such as Convolutional Neural Networks (Kim, 2014; Zhang et al., 2015b), Recurrent Neural Networks (Liu et al., 2016a; Peters et al., 2018) and Transformers (Vaswani et al., 2017; Devlin et al., 2019) have been proposed to effectively capture the semantic properties of textual information by taking the terms' positions into account. This could be interesting in the future work where we would explore the effectiveness of these advanced deep learning techniques on user-venue rating prediction.

Sampling Negative Instances from Implicit Feedback: In Chapter 4, we proposed the Personalised Ranking Framework with Multiple sampling Criteria (PRFMC), which exploits probabilistic models (i.e. the Multi-centre Gaussian model (MGM) (Cheng et al., 2012) and the Social Power-Law Distribution mode (Zhang and Chow, 2015)) to effectively sample negative instances. It will be interesting to investigate whether we can leverage textual information to effectively sample negative instances. For example, Zhang et al. (2015a) proposed a textual-based probabilistic model that captures the users' preferences based on the textual content of their comments. Indeed, this textual-based probabilistic model can be seamlessly integrated into our proposed PRFMC framework. In addition, in Chapter 5, we proposed the *dynamic* geo-based negative sampling approach that effectively samples negative instances based on the sequential order of users' checkins and the geographical information of venues. In particular, our proposed negative sampling approach uniformly samples venues the user has never visited before, but located nearby to the venues he/she visited. Inspired by the negative sampling approaches used in word embeddings (Mikolov et al., 2013b,a; Mnih and Kavukcuoglu, 2013), a noise contrastive estimation techniques (Gutmann and Hyvärinen, 2010) have been widely used to sample words to learn the semantic properties of textual contents. Instead of uniformly sampling nearby venues as negative instances, it will be interesting to investigate whether we can further enhance the quality of context-aware venue recommendation by applying a noise contrastive estimation technique to effectively sample misclassified venues (i.e. the venues the model believes the users have visited but actually they did not) as negative instances.

Modelling Users' Short-Term Preferences from Sequential Feedback: In Chapters 5 and 6, we explored the effectiveness of Recurrent Neural Networks (RNN) in modelling the users' short-term (*dynamic*) preferences from their sequences of checkins. Beside RNN, various approaches (Liu et al., 2018; Kang and McAuley, 2018; Sun et al., 2019) have been recently proposed to exploit self-attention based sequential models (Vaswani et al., 2017; Devlin et al., 2019) to capture the users' dynamic preferences. It will be interesting to extend a self-attention based sequential model to incorporate the contextual information associated with the sequence of users' checkins to effectively capture the users' *dynamic* preferences. Furthermore, since we demonstrated earlier in Chapter 4 that the textual contents of users' comments are useful in modelling the users' preferences, it would also be interesting to explore the users' *dynamic* preferences from the sequence of the users' textual comments.

Cross-Domain Venue Recommendation: In this thesis, we leveraged additional information within the individual LBSN to enhance the effectiveness of context-aware venue recommendation. With the emergence of cross-domain recommendation systems, various Cross-Domain Collaborative Filtering (CDCF) techniques have been proposed to leverage additional information from different domains (Li et al., 2009; Zang and Hu, 2017; Shu et al., 2018; Farseev et al., 2017; Hu et al., 2013). Recently, Manotumruksa et al. (2019) have showed that CDCF models are promising for context-aware venue recommendations. It will be interesting to investigate whether we can exploit those cross-domain techniques to further enhance the effectiveness of our proposed Context Recurrent Collaborative Filtering framework (CRCF). Furthermore, it will be interesting to study whether we can leverage contextual information from different domains to enhance the quality of venue recommendation.

8.3 Closing Remarks

In this thesis, we have addressed a challenging task, namely Context-Aware Venue Recommendation (CAVR). Suggesting interesting venues to the users is a challenging task for a number of reasons. For example, unlike traditional recommendation systems (e.g. movie and book recommendations), the users' preferences on venues change over time (i.e. short- and long-term preferences) and can depend on the current user's context (e.g. the users' current location and time of the day) (Zhang and Chow, 2015; Zhang et al., 2015a). Furthermore, unlike explicit feedback such as the user-venue ratings and comments, the negative feedback cannot be capture from the users' implicit feedback such as their checkins. The lack of negative feedback can hinder the effectiveness of collaborative filtering-based approaches (Rendle et al., 2009). Another challenge of venue recommendation is the problem of cold-start users (i.e. users who have typically only visited and checked in a very small number of all venues in the LBSNs).

We have argued that effective context-aware venue recommendations can be generated by leveraging additional information such as the geographical information of venues, the sequential order of users' checkins and the contextual information associated with the successive checkins (e.g. the time interval and distance between two checkins). To achieve this, we proposed a novel Contextual Recurrent Collaborative Filtering framework (CRCF), which exploits deep neural network architectures to effectively model the users' short-term and long-term contextual preferences from their sequence of checkins. Throughout our comprehensive empirical studies on three large-scale LBSNs, we showed that our proposed CRCF framework can generate more effective venue recommendations than existing venue recommendation systems. Furthermore, to address the challenge of the users' implicit feedback and alleviate the cold-start problem, we proposed a novel negative sampling approach that can effectively sample negative instances by incorporating the geographical information of venues and the users' sequential order of checkins. Our experimental results demonstrated that our proposed negative sampling approach can effectively alleviate the cold-start problems.

We have made progress in addressing some of the main challenges of context-aware venue recommendation. However, there are many interesting and challenging tasks that need to be addressed in order to generate high quality venue recommendations, which we highlighted in Section 8.2. In our various discussions throughout the course of this thesis, it has become apparent that deep neural network techniques are effective in capturing the users' preferences and the characteristics of venues from the users' explicit and implicit feedbacks and alleviating the cold-start problem. We argue that this will continue to be an increasingly important trend in future research on venue recommendation systems.

Appendix A

Summary of Limitations of Previous Works

Table A.1 provide a summary of existing works, their limitations, identified in Chapter 3 as well as technical chapters in this thesis that aim to address these limitations. The table groups existing works into three different groups: namely rating prediction-, top-K venue recommendation- and context-aware venue recommendation-based approaches. The description of the limitations of previous works are summarised below:

Limitation M1: There is a disadvantage in SoReg in that it still relies on the Pearson Correlation Coefficient to estimate the similarity between users.

Limitation M2: There is a disadvantage in CMF where the dimensions of latent factors of venues and comment's terms are similar.

Limitation M3: There is a disadvantage in JMF for jointly learning a user's preference and the characteristics of a venue from a single comment.

Limitation M4: There is a disadvantage in the CMF and JMF models, which treat different latent factors dependently, although these latent factors capture different aspects.

Limitation N1: There is an disadvantage in the NeuMF framework for identifying the top-ranked venues to present to users as it focuses on rating prediction.

Limitation N1: There is an disadvantage in the NeuMF framework for identifying the top-ranked venues to present to users as it focuses on rating prediction.

Limitation N2: MF-based approaches for which this limitation applies (GMF, MLP, NeuMF (He et al., 2017)) assume that the users' preferences are *static* and do not account for the sequential properties of observed feedback.

Table A.1: Summary of existing works and their limitations

Rating prediction-based approaches					
Model	Additional Info	Context	Sequential	Limitations	Chapter
SoReg	users' social links	×	×	M1	4
CMF	comments	×	×	M2-M4	4
JMF	comments	×	×	M2-M4	4
NeuMF	×	×	×	N1, N4, S3	5
GMF	×	×	×	N2-N3	5
MLP	×	×	×	N2-N3	5
Top-K venue recommendation-based approaches					
Model	Additional info	Context	Sequential	Limitations	Chapter
GBPR	venues' location	×	×	S1-S2	4
SBPR	users' social links	×	×	S1-S2	4
SWBPR	users' social links	×	×	S1-S2	4
RNN-MF	×	×	✓	R1-R3	5
DREAM	×	×	✓	R2-R3	5
Context-aware venue recommendation-based approaches					
Model	Additional info	Context	Sequential	Limitations	Chapter
TimeGRU	×	only time	✓	G1-G4	6
CGRU	×	✓	✓	G3-G4	6
LatentCross	×	✓	✓	G3-G4	6

Limitation N3: The MF-based approaches for which this limitation applies (GMF, MLP, NeuMF (He et al., 2017)) ignore the dot product of latent factors that capture user-venue interactions.

Limitation N4: There is an disadvantage in the NeuMF framework that applies the traditional BPR negative sampling approach, in which the contextual information of observed feedback are ignored by the negative sampling approach.

Limitation S1: The sampling approaches for which this limitation applies (GBPR (Yuan et al., 2016), SBPR (Zhao et al., 2014) and SWBPR (Wang et al., 2016)) are built upon predefined sampling assumptions and are not sufficiently flexible to incorporate different types of additional information.

Limitation S2: The sampling approaches for which this limitation applies (GBPR (Yuan et al., 2016), SBPR (Zhao et al., 2014) and SWBPR (Wang et al., 2016)) is based on predefined assumptions, which are contradicted to the previous studies (Cheng et al., 2012; Zhang and Chow, 2015; Zhang et al., 2015a) that examine users' geographical movements and social influences in LBSNs.

Limitation S3: The sampling approaches for which this limitation applies (GBPR (Yuan et al., 2016), SBPR (Zhao et al., 2014) and SWBPR (Wang et al., 2016)) do not take the sequential order of checkins into account.

Limitation R1: The RNN-based factorisation approaches for which this limitation applies (RNN-MF (Zhang et al., 2014b)) do not take the users' long-term (*static*) preferences into account.

Limitation R2: The RNN-based factorisation approaches for which this limitation applies (RNN-MF (Zhang et al., 2014b), BiLSTM-MF (Tang et al., 2017) and DREAM (Yu et al., 2016)) still rely on the dot product operation to combine the latent factors of user ϕu and venues ϕv as well as the hidden unit h_τ when predicting a user's checkin.

Limitation R3: There is an disadvantage in the RNN-based factorisation models that model the user's dynamic preferences from sequential order of checkins by leveraging only the sequence of previously visited venues and ignoring the context associated with the checkins.

Limitation G1: The GRU architecture for which this limitation applies (TimeGRU (Zhu et al., 2017)) can only incorporate the transition context (e.g. the time interval between successive checkins) and is not flexible to incorporate the ordinary context (e.g. user's current location).

Limitation G2: The time gating mechanism proposed by Zhu et al. (2017) is not sufficiently flexible to incorporate multiple types of transition contexts associated with the sequence of checkins.

Limitation G3: The GRU architectures for which this limitation applies (CGRU (Smirnova and Vasile, 2017) and LatentCross (Beutel et al., 2018)) treat the ordinary and transition context similarly. As argued in Section 3.6.2.2, these two types of contexts influence the users' preferences differently and should be treated independently.

Limitation G4: There is an disadvantage in the GRU architectures (CGRU (Smirnova and Vasile, 2017) and LatentCross (Beutel et al., 2018)) that rely on the quantised mapping procedures to represent the transition context.

Bibliography

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning - A Textbook*. Springer.

Aggarwal, C. C. et al. (2016). *Recommender systems*. Springer.

Amatriain, X., Pujol, J. M., and Oliver, N. (2009). I like it... i like it not: Evaluating user ratings noise in recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 247–258. Springer.

Bellogín, A., Castells, P., and Cantador, I. (2017). Statistical biases in information retrieval metrics for recommender systems. *Information Retrieval Journal*, 20(6):606–634.

Beutel, A., Covington, P., Jain, S., Xu, C., Li, J., Gatto, V., and Chi, H. (2018). Latent cross: Making use of context in recurrent recommender systems. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 46–54.

Cañamares, R. and Castells, P. (2018). Should i follow the crowd?: A probabilistic analysis of the effectiveness of popularity in recommender systems. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 415–424. ACM.

Chai, T. and Draxler, R. R. (2014). Root mean square error RMSE or mean absolute error MAE?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250.

Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., and Yu, Y. (2012). Collaborative personalized tweet recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 661–670. ACM.

Chen, L., Chen, G., and Wang, F. (2015). Recommender systems based on user reviews: the state of the art. *User Modeling and User-Adapted Interaction*, 25(2):99–154.

Cheng, C., Yang, H., King, I., and Lyu, M. R. (2012). Fused matrix factorization with geographical and social influence in location-based social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17–23.

Cheng, C., Yang, H., Lyu, M. R., and King, I. (2013). Where you like to go next: Successive point-of-interest recommendation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2605–2611.

Cheng, H.-T., Koc, L., Harmsen, J., , et al. (2016). Wide & deep learning for recommender systems. In *Proceedings of the Workshop on Deep Learning for Recommender Systems*, pages 7–10.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in neural information processing systems*.

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., and Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the Special Interest Group on Computer–Human Interaction*, pages 585–592. ACM.

Cui, B., Tung, A. K., Zhang, C., and Zhao, Z. (2010). Multiple feature fusion for social media applications. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 435–446.

Deshpande, M. and Karypis, G. (2004). Item-based top-n recommendation algorithms. *Proceedings of the ACM Transactions on Information Systems*, 22(1):143–177.

Deveaud, R., Albakour, M.-D., Manotumruksa, J., Macdonald, C., Ounis, I., et al. (2014). Smartvenues: Recommending popular and personalised venues in a city. In *Proceedings of the ACM International Conference on Conference on Information and Knowledge Management*, pages 2078–2080.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Association for Computational Linguistics*.

Dey, A. K., Abowd, G. D., and Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-computer interaction*, 16(2):97–166.

Dinçer, B. T., Macdonald, C., and Ounis, I. (2014). Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 23–32. ACM.

Donkers, T., Loepp, B., and Ziegler, J. (2017). Sequential user-based recurrent neural network recommendations. In *Proceedings of the ACM Conference on Recommender Systems*.

Elkahky, A. M., Song, Y., and He, X. (2015). A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the international conference on World Wide Web*, pages 278–288. International World Wide Web Conferences Steering Committee.

Eravci, B., Bulut, N., Etemoglu, C., and Ferhatosmanoğlu, H. (2016). Location recommendations for new businesses using check-in data. In *Proceedings of the International Conference on Data MiningW*, pages 1110–1117. IEEE.

Farseev, A., Samborskii, I., Filchenkov, A., and Chua, T.-S. (2017). Cross-domain recommendation via clustering on multi-layer graphs. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 195–204.

Fu, X. and Li, X. (2015). From movie reviews to restaurants recommendation. Technical report, Stanford University. http://stanford.io/2aB0T0f.

Fuhr, N. (2018). Some common mistakes in IR evaluation, and how they can be avoided. In *ACM SIGIR Forum*, volume 51, pages 32–41. ACM.

Gao, H., Tang, J., Hu, X., and Liu, H. (2013). Exploring temporal effects for location recommendation on location-based social networks. In *Proceedings of the ACM Conference on Recommender Systems*, pages 93–100.

Garcin, F., Faltings, B., Donatsch, O., Alazzawi, A., Bruttin, C., and Huber, A. (2014). Offline and online evaluation of news recommender systems at swissinfo. ch. In *Proceedings of the ACM Conference on Recommender Systems*, pages 169–176. ACM.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 315–323.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232.

Griesner, J.-B., Abdessalem, T., and Naacke, H. (2015). Poi recommendation: Towards fused matrix factorization with geographical and temporal influences. In *Proceedings of the ACM Conference on Recommender Systems*, pages 301–304. ACM.

Guo, G., Zhang, J., Sun, Z., and Yorke-Smith, N. (2015a). LibRec: A Java library for recommender systems. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*.

Guo, G., Zhang, J., and Yorke-Smith, N. (2015b). TrustSVD: Collaborative filtering with both the explicit and implicit influence of user trust and of item ratings. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 297–304.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 770–778.

He, R. and McAuley, J. (2016). VBPR: Visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 144–150.

He, X. and Chua, T.-S. (2017). Neural factorization machines for sparse predictive analytics. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 355–364.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. (2017). Neural collaborative filtering. In *Proceedings of the International Conference on World Wide Web*, pages 173–182.

He, X., Zhang, H., Kan, M.-Y., and Chua, T.-S. (2016b). Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 549–558.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hu, L., Cao, J., Xu, G., Cao, L., Gu, Z., and Zhu, C. (2013). Personalized recommendation via cross-domain triadic factorization. In *Proceedings of the international conference on World Wide Web*, pages 595–606.

Hu, L., Sun, A., and Liu, Y. (2014). Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 345–354.

Hu, Y., Koren, Y., and Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Proceedings of the International Conference on Data Mining*, pages 263–272.

Huang, J., Rogers, S., and Joo, E. (2014). Improving restaurants by extracting subtopics from Yelp reviews. *iConference (Social Media Expo)*.

Hurley, N. and Zhang, M. (2011). Novelty and diversity in top-n recommendation–analysis and evaluation. *Proceedings of the ACM Transactions on Internet Technology*, 10(4):14.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *Proceedings of the ACM Transactions on Information Systems*, 20(4):422–446.

Jin, Z., Li, Q., Zeng, D. D., Zhan, Y., Liu, R., Wang, L., and Ma, H. (2016). Jointly modeling review content and aspect ratings for review rating prediction. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 893–896.

Jing, H. and Smola, A. J. (2017). Neural survival recommender. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 515–524.

Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *Proceedings of the International Conference on International Conference on Machine Learning*, pages 2342–2350.

Kang, W.-C. and McAuley, J. (2018). Self-attentive sequential recommendation. In *Proceedings of the International Conference on Data Mining)*, pages 197–206. IEEE.

Kharitonov, E. (2016). *Using interaction data for improving the offline and online evaluation of search engines*. PhD thesis, University of Glasgow.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181.

Koren, Y. (2010a). Collaborative filtering with temporal dynamics. volume 53, pages 89–97. ACM.

Koren, Y. (2010b). Factor in the neighbors: Scalable and accurate collaborative filtering. *Transactions on Knowledge Discovery from Data (TKDD)*, 4(1):1.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, (8):30–37.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

Lee, J. Y. and Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Lee, S., Song, S.-i., Kahng, M., Lee, D., and Lee, S.-g. (2011). Random walk based entity ranking on graph for multidimensional recommendation. In *Proceedings of the ACM Conference on Recommender Systems*, pages 93–100. ACM.

Li, B., Yang, Q., and Xue, X. (2009). Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Li, H., Ge, Y., and Zhu, H. (2016). Point-of-interest recommendations: Learning potential check-ins from friends. In *Proceedings of the Knowledge Discovery and Data Mining*, pages 975–984.

Li, X., Cong, G., Li, X.-L., Pham, T.-A. N., and Krishnaswamy, S. (2015). Rank-GeoFM: A ranking based geographical factorization method for point of interest recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 433–442.

Lian, D., Zhao, C., Xie, X., Sun, G., Chen, E., and Rui, Y. (2014). GeoMf: Joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 831–840.

Linden, G., Smith, B., and York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, (1):76–80.

Liu, B., Fu, Y., Yao, Z., and Xiong, H. (2013). Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1043–1051. ACM.

Liu, P., Qiu, X., and Huang, X. (2016a). Recurrent neural network for text classification with multi-task learning. *Proceedings of the International Joint Conference on Artificial Intelligence*.

Liu, Q., Wu, S., Wang, D., Li, Z., and Wang, L. (2016b). Context-aware sequential recommendation. In *Proceedings of the International Conference on Data Mining*, pages 1053–1058.

Liu, Q., Wu, S., Wang, L., and Tan, T. (2016c). Predicting the next location: A recurrent model with spatial and temporal contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 152–160.

Liu, Q., Zeng, Y., Mokhosi, R., and Zhang, H. (2018). STAMP: Short-term attention/memory priority model for session-based recommendation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1831–1839. ACM.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, pages 225–331.

Loni, B., Pagano, R., Larson, M., and Hanjalic, A. (2016). Bayesian personalized ranking with multi-channel user feedback. In *Proceedings of the ACM Conference on Recommender Systems*, pages 361–364.

Lu, Q., Chen, T., Zhang, W., Yang, D., and Yu, Y. (2012). Serendipitous personalized ranking for top-n recommendation. In *Proceedings of the ACM International Conference on Web Intelligence*, volume 1, pages 258–265. IEEE.

Ma, H. (2014). On measuring social friend interest similarities in recommender systems. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 465–474.

Ma, H., Zhou, D., Liu, C., Lyu, M. R., and King, I. (2011). Recommender systems with social regularization. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 287–296.

Manotumruksa, J. (2017). Deep collaborative filtering approaches for context-aware venue recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1383–1383. ACM.

Manotumruksa, J., Macdonald, C., and Ounis, I. (2016). Regularising factorised models for venue recommendation using friends and their comments. In *Proceedings of the ACM International Conference on Conference on Information and Knowledge Management*, pages 1981–1984.

Manotumruksa, J., Macdonald, C., and Ounis, I. (2017a). A deep recurrent collaborative filtering framework for venue recommendation. In *Proceedings of the ACM International Conference on Conference on Information and Knowledge Management*, pages 1429–1438.

Manotumruksa, J., Macdonald, C., and Ounis, I. (2017b). A personalised ranking framework with multiple sampling criteria for venue recommendation. In *Proceedings of the ACM International Conference on Conference on Information and Knowledge Management*, pages 1469–1478.

Manotumruksa, J., Macdonald, C., and Ounis, I. (2018). A contextual attention recurrent architecture for context-aware venue recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 555–564. ACM.

Manotumruksa, J., Rafailidis, D., Macdonald, C., and Ounis, I. (2019). On cross-domain transfer in venue recommendation. In *Proceedings of the European Conference on Information Retrieval*, pages 443–456. Springer.

McDonald, G., Macdonald, C., and Ounis, I. (2017). Enhancing sensitivity classification with semantic features using word embeddings. In *Proceedings of the European Conference on Information Retrieval*, pages 450–463. Springer.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., and Riedl, J. (2003). Movielens unplugged: experiences with an occasionally connected recommender system. In *Proceedings of the ACM International Conference on Intelligent User Interfaces*, pages 263–266. ACM.

Mnih, A. and Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, pages 2265–2273. Curran Associates, Inc.

Musto, C., Semeraro, G., De Gemmis, M., and Lops, P. (2015). Word embedding techniques for content-based recommender systems: an empirical evaluation. In *Proceedings of the ACM Conference on Recommender Systems*.

Musto, C., Semeraro, G., de Gemmis, M., and Lops, P. (2016). Learning word embeddings from Wikipedia for content-based recommender systems. In *Proceedings of the European Conference on Information Retrieval*, pages 729–734.

Narayanan, M. and Cherukuri, A. K. (2016). A study and analysis of recommendation systems for location-based social network (lbsn) with big data. *IIMB Management Review*, 28(1):25–30.

Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011). An empirical study of geographic user activity patterns in Foursquare. *Proceedings of the Interational AAAI Conference on Web and Social Media*, pages 25–30.

Ozsoy, M. G. (2016). From word embeddings to item recommendation. *arXiv:1601.01356*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the Association for Computational Linguistics*.

Preoţiuc-Pietro, D. and Cohn, T. (2013). Mining user behaviours: a study of check-in patterns in location based social networks. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 306–315. ACM.

Pu, P. and Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20(6):542–556.

Rendle, S. (2012). Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, page 57.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Association for Uncertainty in Artificial Intelligence*, pages 452–461.

Rendle, S., Freudenthaler, C., and Schmidt-Thieme, L. (2010). Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the International Conference on World Wide Web*, pages 811–820.

Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3):56–58.

Ricci, F., Rokach, L., and Shapira, B. (2015a). Evaluating recommendation systems. In *Recommender Systems Handbook*, pages 257–294. Springer.

Ricci, F., Rokach, L., and Shapira, B. (2015b). Recommender systems: introduction and challenges. In *Recommender Systems Handbook*, pages 1–34. Springer.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, 5(3):1.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *In Proceedings of the International Conference on World Wide Web*, pages 285–295. ACM.

Shani, G., Heckerman, D., and Brafman, R. I. (2005). An MDP-based recommender system. *Journal of Machine Learning Research*, 6(Sep):1265–1295.

Sheugh, L. and Alizadeh, S. H. (2015). A note on Pearson correlation coefficient as a metric of similarity in recommender system. In *AI & Robotics (IRANOPEN), 2015*, pages 1–6. IEEE.

Shi, Y., Karatzoglou, A., Baltrunas, L., Larson, M., Oliver, N., and Hanjalic, A. (2012). CLiMF: learning to maximize reciprocal rank with collaborative less-is-more filtering. In *Proceedings of the ACM Conference on Recommender Systems*, pages 139–146.

Shu, K., Wang, S., Tang, J., Wang, Y., and Liu, H. (2018). Crossfire: Cross media joint friend and item recommendations. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 522–530.

Sinha, R. R., Swearingen, K., et al. (2001). Comparing recommendations made by online systems and friends. In *DELOS*.

Skansi, S. (2018). *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*. Springer.

Smirnova, E. and Vasile, F. (2017). Contextual sequence modeling for recommendation with recurrent neural networks. In *Proceedings of the Workshop on Deep Learning for Recommender Systems*, pages 2–9.

Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2222–2230.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., and Jiang, P. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. volume abs/1904.06690.

Tan, Y. K., Xu, X., and Liu, Y. (2016). Improved recurrent neural networks for session-based recommendations. In *Proceedings of the Workshop on Deep Learning for Recommender Systems*, pages 17–22.

Tang, D., Qin, B., Liu, T., and Yang, Y. (2015). User modeling with neural network for review rating prediction. In *Proceedings of the International Conference on World Wide Web*, pages 1340–1346.

Tang, S., Wu, Z., and Chen, K. (2017). Movie recommendation via BLSTM. In *Proceedings of the International Conference on Multimedia Modeling*, pages 269–279.

Teow, M. Y. (2017). Understanding convolutional neural networks using a minimal model for handwritten digit recognition. In *Proceedings of the International Conference on Automatic Control and Intelligent Systems*, pages 167–172. IEEE.

Vargas, S. (2015). *Novelty and diversity evaluation and enhancement in recommender systems*. PhD thesis, Ph. D. thesis.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Voorhees, E. M. et al. (1999). The trec-8 question answering track report. In *Proceedings of the Text Retreival Conference*, volume 99, pages 77–82.

Voorhees, E. M. and Harman, D. (2003). Common evaluation measures. In *Proceedings of the Text Retreival Conference*, pages 500–255.

Wang, L., Bennett, P. N., and Collins-Thompson, K. (2012). Robust ranking models via risk-sensitive optimization. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 761–770. ACM.

Wang, X., Lu, W., Ester, M., Wang, C., and Chen, C. (2016). Social recommendation with strong and weak ties. In *Proceedings of the ACM International Conference on Conference on Information and Knowledge Management*, pages 5–14.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30(1):79–82.

Wu, Y., DuBois, C., Zheng, A. X., and Ester, M. (2016). Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 153–162. ACM.

Xiang, L., Yuan, Q., Zhao, S., Chen, L., Zhang, X., Yang, Q., and Sun, J. (2010). Temporal recommendation on graphs via long-and short-term preference fusion. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 723–732. ACM.

Yang, D., Zhang, D., Zheng, V. W., and Yu, Z. (2015). Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *Systems, Man, and Cybernetics: Systems, IEEE Transactions on*, 45(1):129–142.

Yang, X., Macdonald, C., and Ounis, I. (2018). Using word embeddings in Twitter election classification. *Information Retrieval Journal*, 21(2-3):183–207.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the Association for Computational Linguistics*, pages 1480–1489.

Yao, Z., Fu, Y., Liu, B., Liu, Y., and Xiong, H. (2016). POI recommendation: A temporal matching between POI popularity and user regularity. In *Proceedings of the International Conference on Data Mining*, pages 549–558. IEEE.

Ye, M., Yin, P., Lee, W.-C., and Lee, D.-L. (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 325–334.

Ying, H., Chen, L., Xiong, Y., and Wu, J. (2016). Pgrank: Personalized geographical ranking for point-of-interest recommendation. In *Proceedings of the International Conference on World Wide Web*, pages 137–138.

Ying, J. J.-C., Lu, E. H.-C., Kuo, W.-N., and Tseng, V. S. (2012). Urban point-of-interest recommendation by mining user check-in behaviors. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 63–70. ACM.

Yu, F., Liu, Q., Wu, S., Wang, L., and Tan, T. (2016). A dynamic recurrent model for next basket recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 729–732.

Yu, Y., Tang, S., Zimmermann, R., and Aizawa, K. (2014). Empirical observation of user activities: Check-ins, venue photos and tips in Foursquare. In *Proceedings of the ACM International Workshop on Internet-Scale Multimedia Management*, pages 31–34. ACM.

Yuan, F., Guo, G., Jose, J., Chen, L., and Yu, H. (2016). Joint geo-spatial preference and pairwise ranking for point-of-interest recommendation. In *Proceedings of the International Conference on Tools with Artificial Intelligence*, pages 46–53.

Yuan, Q., Cong, G., Ma, Z., Sun, A., and Thalmann, N. M. (2013). Time-aware point-of-interest recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372.

Zang, Y. and Hu, X. (2017). LKT-FM: A novel rating pattern transfer model for improving non-overlapping cross-domain collaborative filtering. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 641–656. Springer.

Zhang, H., Yang, Y., Luan, H., Yang, S., and Chua, T.-S. (2014a). Start from scratch: Towards automatically identifying, modeling, and naming visual attributes. In *Proceedings of the ACM International Conference on Multimedia*, pages 187–196.

Zhang, J.-D. and Chow, C.-Y. (2013). iGSLR: personalized geo-social location recommendation: a kernel density estimation approach. In *Proceedings of the International Conference on Advances in Geographic Information Systems*, pages 334–343.

Zhang, J.-D. and Chow, C.-Y. (2015). GeoSoCa: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 443–452.

Zhang, J.-D., Chow, C.-Y., and Zheng, Y. (2015a). ORec: An opinion-based point-of-interest recommendation framework. In *Proceedings of the ACM International Conference on Conference on Information and Knowledge Management*, pages 1641–1650.

Zhang, X., Zhao, J., and LeCun, Y. (2015b). Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657.

Zhang, Y., Dai, H., Xu, C., Feng, J., Wang, T., Bian, J., Wang, B., and Liu, T.-Y. (2014b). Sequential click prediction for sponsored search with recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1369–1375.

Zhang, Y., Roller, S., and Wallace, B. (2016). MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification. *Proceedings of the Association for Computational Linguistics*.

Zhao, S., Zhao, T., Yang, H., Lyu, M. R., and King, I. (2016). STELLAR: Spatial-temporal latent ranking for successive point-of-interest recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhao, T., McAuley, J., and King, I. (2014). Leveraging social connections to improve personalized ranking for collaborative filtering. In *Proceedings of the ACM International Conference on Conference on Information and Knowledge Management*, pages 261–270.

Zheng, L., Noroozi, V., and Yu, P. S. (2017). Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 425–434. ACM.

Zhou, C., Sun, C., Liu, Z., and Lau, F. (2015). A C-LSTM neural network for text classification. *In Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., and Xu, B. (2016). Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *Proceedings of the International Committee on Computational Linguistics*.

Zhu, Y., Li, H., Liao, Y., Wang, B., Guan, Z., Liu, H., and Cai, D. (2017). What to do next: Modeling user behaviors by Time-LSTM. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3602–3608.